DOCUMENT RESUME

ED 198 164 TM 810 131

AUTHOR	Weiss, David J., Ed.
TITLE	Proceedings of the 1979 Computerized Adaptive Testing Conference (Wayzata, Minnesota, June 27-30, 1979).
INSTITUTION	Minnesota Univ., Minneapolis. Dept. of Psychology.
SPONS AGENCY	Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE	Sep 80
CONTRACT	N00014-79-C-0196
NOTE	467p.: This conference and the preparation of these Proceedings were jointly sponsored by the Office of Naval Research, Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, Military Enlistment Processing Command, and the Navy Personnel Research and Development Center.
EDRS PRICE	MF01/PC19 Plus Postage.
DESCRIPTORS	Academic Ability: Academic Achievement: *Comparative Testing: *Computer Assisted Testing: *Evaluation Methods: Foreign Countries: *Latent Trait Theory: Longitudinal Studies: Mastery Tests: Mathematical Models: Nonparametric Statistics: Reaction Time: Test Reliability: Test Validity
IDENTIFIERS	Germany: *Rasch Model

ABSTRACT

This report is the Proceedings of the third conference of its type. Included are 23 of the 25 papers presented at the conference, discussion of these papers by invited discussants, and symposium papers by a group of leaders in adaptive testing and latent trait test theory research and applications. The papers are organized into the following topical sessions: (1) Adaptive Testing Strategies for Measuring Ability: (2) Adaptive Testing in Germany: (3) Adaptive Mastery Testing: (4) Estimating Response Functions without Assuming a Parametric Model: (5) Item Linking and Equating: (6) Latent Trait Models which Incorporate Response Time as well as Response Appropriateness: (7) Person-Item Interaction: (8) Use of Latent Trait Models with Estimated Item Parameters: and (9) Longitudinal Measurement with Latent Trait Models. A symposium and discussion on the State of the Art of Adaptive Testing and Latent Trait Test Theory included presentations by Gerhard Fischer, Frederic M. Lord, James Lumsden, and David J. Weiss. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED198164

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM | IR

In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

PROCEEDINGS OF THE

1979 COMPUTERIZED ADAPTIVE TESTING CONFERENCE

EDITED BY
DAVID J. WEISS

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Office of Naval Research
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 131

Prepared under contract No. N00014-79-C-0196, NR150-432
with the Personnel and Training Research Programs
Psychological Sciences Division
Office of Naval Research

Approved for public release; distribution unlimited
Reproduction in whole or in part is permitted for
any purpose of the United States Government.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Proceedings of the
1979 Computerized Adaptive Testing Conference

5. TYPE OF REPORT & PERIOD COVERED
Conference Proceedings
Report, 27-30 June 1979

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Edited by David J. Weiss

8. CONTRACT OR GRANT NUMBER(s)
N00014-79-C-0196

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Psychology
University of Minnesota
Minneapolis, MN 55455

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
P.E.: 61153N Proj: RR042-04
T.A.: RR042-04-01
W.U.: NR150-432

11. CONTROLLING OFFICE NAME AND ADDRESS
Personnel and Training Research Programs
Office of Naval Research
Arlington, VA 22217

12. REPORT DATE
September 1980

13. NUMBER OF PAGES
454

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution limited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This conference and the preparation of these Proceedings were jointly sponsored by the Office of Naval Research, Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, Military Enlistment Processing Command, and the Navy Personnel Research and Development Center.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

ability testing
achievement testing
sequential testing
automated testing
branched testing
individualized testing
tailored testing
programmed testing
response-contingent testing
item characteristic curve theory
latent trait test theory
test theory
item response theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is the Proceedings of the 1979 Computerized Adaptive Testing Conference held June 27 to 30, 1979, at the Spring Hill Conference Center in Wayzata, Minnesota. These Proceedings include 23 of the 25 papers presented at the conference, discussion of these papers by invited discussants, and symposium papers by a group of leaders in adaptive testing and latent trait test theory research and applications.

(continued on the other side)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

The papers are organized into the following topical sessions:

1. Adaptive Testing Strategies for Measuring Ability. Papers by James R. McBride; Marilyn F. Johnson and David J. Weiss; and Steven Gorman. Discussion by Brian Waters.

2. Adaptive Testing in Germany. Papers by Lutz F. Hornke and Michael B. Sauter; and Wolfgang Wildegrube.

3. Adaptive Mastery Testing. Papers by Mark Reckase; Stanley Kalisch; and G. Gage Kingsbury and David J. Weiss. Discussion by Melvin Novick.

4. Estimating Response Functions without Assuming a Parametric Model. Paper by Fumiko Samejima. Discussion by Robert Tsutakawa.

5. Item Linking and Equating. Papers by Gary Marco, Nancy Petersen, and Elizabeth Stewart; Wendy Yen; and Malcolm J. Ree and Harald E. Jensen. Discussion by Gail Ironson.

6. Latent Trait Models which Incorporate Response Time as well as Response Appropriateness. Papers by Kikumi and Maurice Tatsuoka; and David Thissen. Discussion by John B. Carroll.

7. Person-Item Interaction. Papers by Ronald Mead; Howard Wainer and Benjamin Wright; and Michael Levine and Fritz Drasgow. Discussion by James Lumsden.

8. Use of Latent Trait Models with Estimated Item Parameters. Papers by Ronald K. Hambleton and Linda L. Cook; Michael Waller; H. Swaminathan and Janice Gifford; and Frederic M. Lord. Discussion by Bert F. Green, Jr.

9. Longitudinal Measurement with Latent Trait Models. Papers by Gerhard Fischer; R. Darrell Bock; and Lalitha Sanathanan.

Symposium and Discussion: State of the Art of Adaptive Testing and Latent Trait Test Theory. Presentations by Gerhard Fischer; Frederic M. Lord; James Lumsden; and David J. Weiss. Comment by John B. Carroll.

PROCEEDINGS OF THE
1979
COMPUTERIZED ADAPTIVE TESTING
CONFERENCE

held at the
Spring Hill Conference Center, Wayzata, Minnesota
June 27-30, 1979

EDITED BY
DAVID J. WEISS

COMPUTERIZED ADAPTIVE TESTING LABORATORY
PSYCHOMETRIC METHODS PROGRAM
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF MINNESOTA
SEPTEMBER 1980

Prepared under contract No. N00014-79-C-0196, NR150-432
with the Personnel and Training Research Programs,
Psychological Sciences Division, Office of Naval Research

Approved for public release, distribution unlimited.
Reproduction in whole or in part is permitted for
any purpose of the United States Government

ACKNOWLEDGMENTS

THE 1979 COMPUTERIZED ADAPTIVE TESTING CONFERENCE WAS JOINTLY SPONSORED BY THE OFFICE OF NAVAL RESEARCH, THE NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER, THE AIR FORCE OFFICE OF SCIENTIFIC RESEARCH, THE ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES, THE MILITARY ENLISTMENT PROCESSING COMMAND, AND THE DEFENSE ADVANCED RESEARCH PROJECTS AGENCY. THE CONFERENCE WAS HELD AT THE SPRING HILL CONFERENCE CENTER, WAYZATA, MINNESOTA, JUNE 27 TO 30, 1979. THE CONTRIBUTIONS OF THE FOLLOWING INDIVIDUALS TO THE CONFERENCE AND THE PREPARATION OF THESE PROCEEDINGS ARE GRATEFULLY ACKNOWLEDGED.

CONFERENCE PLANNING COMMITTEE	CHARLES DAVIS DAVID J. WEISS
TECHNICAL EDITOR	BARBARA L. CAMM
ARTIST	DONALD BLAKESLEE
PHOTOGRAPHER	JOHN MARTIN
TYPISTS	LAURA BERQUAM JAMES DEGERSTROM MELINDA HUTSON ROXANNE KEENE KRISTI LINDENBERG DIANE SWEENEY LYNNE WEYENBERG-DARKOW
WORD PROCESSING	PROGRAM TYPST WRITTEN BY KAREN SCHAFFER

CONTENTS

Conference Introduction....................................Marshall J. Farr 1

Session 1: Adaptive Testing Strategies for Measuring Ability 3

Adaptive Verbal Ability Testing in a Military Setting.....James R. McBride 4
Parallel Forms Reliability and Measurement Accuracy Comparison of Adaptive and Conventional Testing Strategies.................Marilyn F. Johnson and David J. Weiss 16
A Comparison of the Accuracy of Bayesian Adaptive and Static Tests Using a Correction for Regression.........Steven Gorman 35
Discussion..Brian Waters 51

Session 2: Adaptive Testing in Germany 56

A Validity Study of an Adaptive Test of Reading Comprehension............Lutz F. Hornke and Michael P. Sauter 57
Computerized Testing in the Federal Armed Forces..........................Wolfgang Wildgrube 68

Session 3: Adaptive Mastery Testing 78

Some Decision Procedures for Use with Tailored Testing...................................Mark D. Reckase 79
A Model for Computerized Adaptive Testing Related to Instructional Situations..................Stanley J. Kalisch 101
A Comparison of ICC-Based Adaptive Mastery Testing and the Waldian Probability Ratio Method...........G. Gage Kingsbury and David J. Weiss 120
Discussion...Melvin Novick 140

Session 4: Estimating Response Functions Without Assuming a Parametric Model 144

Constant Information Model on the Dichotomous Response Level...............................Fumiko Samejima 145
Discussion...Robert Tsutakawa 164

Session 5: Item Linking and Equating 166

A Test of the Adequacy of Curvilinear Score Equating Models........Gary Marco, Nancy Petersen, and Elizabeth Stewart 167
The Effects of Context on Latent Trait Model Item Parameter and Trait Estimates........................Wendy M. Yen 197
Effects of Sample Size on Linear Equating of Item Characteristic Curve Parameters....................Malcolm J. Ree and Harald E. Jensen 218
Discussion..Gail Ironson 229

Session 6: Latent Trait Models Which Incorporate Response Time as Well as Response Appropriateness 235

A Model for Incorporating Response-Time Data in Scoring Achievement Tests...Kikumi Tatsuoka and Maurice Tatsuoka 236
Latent Trait Scoring of Timed Ability Tests..................David Thissen 257
Discussion..John B. Carroll 278

Session 7: Person-Item Interaction 284

Using the Rasch Model to Identify Person-Based Measurement Disturbances....................Ronald J. Mead 285
Robust Estimation in the Rasch Model.................Howard Wainer and Benjamin D. Wright 301
Appropriateness Measurement: Basic Principles and Validating Studies.................Michael Levine and Fritz Drasgow 322
Discussion...James Lumsden 345

Session 8: Use of Latent Trait Models with Estimated Item Parameters 348

Robustness of Latent Trait Models and Effects of Test Length and Sample Size on the Precision of Ability Estimates.....Ronald K. Hambleton and Linda L. Cook 349
Estimating Abilities Within the Two-Parameter Logistic Latent Trait Model in the Presence of A Non-Symmetric Distribution of Ability..................Michael Waller 365
Estimation of Parameters in the Latent Trait Model..........Hariharan Swaminathan and Janice Gifford 372
Small N Justifies Rasch Methods..........................Frederic M. Lord 386
Discussion...Bert F. Green, Jr. 396

Session 9: Longitudinal Measurement with Latent Trait Models 399

Some Latent Trait Models for Measuring Change in Qualitative Observations.............................Gerhard Fischer 400
The Mental Growth Curve Re-Examined........................R. Darrell Bock 415
Latent Structure Estimation for Assessing Gain in Ability.......................Lalitha Sanathanan 425

Symposium and Discussion: State of the Art of Adaptive Testing and Latent Trait Test Theory 435

Gerhard Fischer.. 436
Frederic M. Lord... 439
James Lumsden.. 442
David J. Weiss... 444
Comment: John B. Carroll... 449

Appendix: Addresses of Conference Participants......................... 453

CONFERENCE INTRODUCTION

MARSHALL J. FARR, DIRECTOR
PERSONNEL AND TRAINING RESEARCH PROGRAMS
OFFICE OF NAVAL RESEARCH

I am proud to introduce these Proceedings of the 1979 Computerized Adaptive Testing Conference. This was the third conference of its type sponsored by the Office of Naval Research (ONR) in conjunction with various co-sponsors, which for this conference included the Navy Personnel Research and Development Center, the Air Force Office of Scientific Research, the Army Research Institute for the Behavioral and Social Sciences, the Military Enlistment Processing Command, and the Defense Advanced Research Projects Agency.

The growing international interest in computerized adaptive testing was evidenced by the fact that representatives from Australia, Austria, Belgium, Japan, and West Germany made up part of the more than 80 invited participants in this conference. Equally impressive was the widespread representation from federal agencies: In addition to those from the sponsors, participants came from the U.S. Marine Corps, Air Force Human Resource Laboratories, the U.S. Coast Guard, the Navy Guided Missile School, the U.S. Civil Service Commission, and the Naval Aerospace Medical Research Laboratory.

Computerized adaptive testing (CAT) has come a long way in a short span of years, thanks to an ever-burgeoning interest in the field, which continues to be spearheaded by the ONR contractors represented in these proceedings. Since the 1977 CAT conference, the Defense Department has formally recognized its promise. In January 1979 a memorandum issued at the level of the Office of the Secretary of Defense directed "the development and further evaluation of the feasibility of implementing computerized adaptive testing in the Department of Defense." The memorandum went on to call for a Defense Department-wide research and development program, which will eventually transform the Armed Services Vocational Aptitude Battery (ASVAB)—now used by all the Services for enlisted personnel selection and classification—into a computerized adaptive examination. That implementation-feasibility study is now underway, guided by a steering committee representing the Navy, Army, Air Force, Marine Corps, and the Military Enlistment Processing Command (MEPCOM).

The Office of Naval Research is generally acknowledged as one of the paramount forces, if not the leader, in constructing the theoretical research foundations that make CAT possible. The proceedings of this conference demonstrate that our support of research on important theoretical questions in CAT continues unabated. I believe strongly in the potential of CAT for possibly revolutionizing test administration and scoring in the measurement of both ability and achievement. I further believe that having computers readily available for online testing will encourage the development of new kinds of test items, administration, and scoring, over and above the changes to be wrought by CAT.

SESSION 1*:
ADAPTIVE TESTING STRATEGIES FOR MEASURING ABILITY

ADAPTIVE VERBAL ABILITY TESTING IN A MILITARY SETTING

JAMES R. MCBRIDE
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

PARALLEL FORMS RELIABILITY AND MEASUREMENT ACCURACY COMPARISON OF ADAPTIVE AND CONVENTIONAL TESTING STRATEGIES

MARILYN F. JOHNSON AND DAVID J. WEISS
UNIVERSITY OF MINNESOTA

A COMPARISON OF THE ACCURACY OF BAYESIAN ADAPTIVE AND STATIC TESTS USING A CORRECTION FOR REGRESSION

STEVEN GORMAN
DEPARTMENT OF THE NAVY

DISCUSSION

BRIAN WATERS
AIR UNIVERSITY

---

*A paper entitled "Criterion-Related Validity of Conventional and Adaptive Ability Tests in a Military Environment," by James B. Sympson, was also presented in Session 1, but was not available for inclusion in these Proceedings.

ADAPTIVE VERBAL ABILITY TESTING IN A MILITARY SETTING

JAMES R. MCBRIDE
NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER

Since January 1976 all military services have used a common battery of mental tests for enlisted personnel selection and classification: the Armed Services Vocational Aptitude Battery (ASVAB). The battery includes 12 subtests of cognitive aptitudes. These subtests are necessarily short; they are usually scored by hand; the raw scores are manually converted into service-specific scaled scores using conversion tables; and the scale scores are manually recorded and manually transcribed into permanent individual personnel records.

The U.S. Marine Corps has identified some difficulties with the ASVAB testing program. Now that the ASVAB has supplanted service-specific classification test batteries, a single test battery must serve all the special testing needs of the four services. In many cases, ASVAB subtests are excessively difficult for Marine Corps selection and classification purposes; this can result in inefficient and inaccurate classification. There has been some compromise of ASVAB test security: Test booklets and answer keys have been stolen. This problem, if uncontrolled, could seriously degrade the validity of the tests for classification purposes. The manual nature of the test scoring, score conversion, and score recording procedures provides opportunity for clerical error, and it is believed that such errors may have resulted in numerous accession errors.

The Marine Corps formulated an operational requirement to lessen or eliminate the impact of the problems discussed above. Computer-administered adaptive testing (CAT) was identified as one potential solution to all of these problems. In an adaptive test, test difficulty is tailored dynamically to the ability level of the individual examinee; in principle, then, CAT eliminates the problem of excessive test difficulty and should yield scores that promote accurate selection and classification decisions. CAT addresses the test security problem by eliminating printed booklets and scoring keys and by administering an individually tailored set of test items to each examinee. Additionally, since CAT automates test administration, test scoring and recording are automated as well, thereby eliminating human clerical error from the testing system.

Recognizing the potential of CAT for selection and classification testing, the Marine Corps tasked NPRDC with investigating the feasibility of CAT as part of a program of phased research and development related to military personnel accessioning.

Purpose

The research reported here was intended to assess the feasibility of using computerized adaptive testing (CAT) in a Marine Corps recruit/applicant population and, at the same time, to verify the claimed merits of CAT as a psychological measurement technique. These two research issues could only be addressed by administering adaptive tests to appropriate examinee samples. The capability to do this had to be developed--equipment identified, software written, and large banks of test items assembled and calibrated using item characteristic curve (ICC) models. After this development was completed, a pilot study involving verbal ability tests was conducted. This report describes the pilot study of the feasibility and psychometric merits of an adaptive procedure for measuring verbal ability.

Background

Group-administered paper-and-pencil "objective" ability tests date back to World War I, when the introduction of the Army Alpha test signalled an era of vast improvements in the administrative efficiency of psychological testing. The price paid for this efficiency was loss of flexibility, since all examinees had to answer a common set of test questions. The psychometric effect of this was not too serious, provided that a test was designed to have a difficulty level appropriate to its intended application or that a test was sufficiently long to overcome minor design deficiencies. For persons whose ability level was not near the target difficulty level of the test, however, the paper-and-pencil test was not a particularly accurate or precise measuring instrument.

The psychological tests used by the armed services for selection and classification are group-administered paper-and-pencil tests. Such tests, as just discussed, lack the flexibility to measure well over a wide range of ability. In order to achieve that flexibility, the difficulty level of the test would have to be chosen to fit individual ability levels. Since individual ability levels are not known prior to testing, this is not practical; however, it can be accomplished using an adaptive test in which test items are chosen sequentially on the basis of the examinee's performance. This sequential item choice can best be accomplished using automated test administration, for example, by having the test administered at an interactive computer terminal.

The historical development of computer-administered adaptive testing was reviewed by Weiss and Betz (1973) and by Wood (1973). Weiss surveyed a variety of alternative adaptive testing methods (1974) and summarized a number of potential advantages of CAT over conventional paper-and-pencil tests (1975). Despite those potential advantages, most research into adaptive testing had been at the basic research level, until 1975 when the U.S. Civil Service Commission began moving toward early 1980s implementation of computer-based adaptive administration of its PACE examination (Gorham, 1975).

The U.S. Civil Service Commission's implementation plans were based on research conducted by Urry and his colleagues (e.g., Urry, 1977). Urry chose to adopt a Bayesian sequential adaptive testing procedure proposed by Owen (1969, 1975) and demonstrated that the procedure could achieve satisfactory levels of measurement reliability in substantially less than half the number of items required of a conventional test; in one instance he estimated that an adaptive test was equivalent in reliability to a conventional test five times as long (Urry, 1977). It is this efficiency of measurement which has motivated most psychometric interest in adaptive testing, although test users have often been more attracted by its practical advantages, which were discussed above.

Marine Corps interest in CAT for personnel selection and classification testing resulted from dissatisfaction with certain aspects of the joint service paper-and-pencil testing battery. Subtests used for selection decisions were also used as a basis for personnel classification and assignment to specialized training; a test designed for one of these purposes would likely be inappropriate for the other, and this might result in disproportionate numbers of selection or assignment errors. Clerical errors in the manual scoring and score recording processes were felt to be another serious source of accessioning errors; and the effects of test compromise were inevitable with the use of the same test battery over a period of several years.

Recognizing that computerized test administration could eliminate scoring and clerical errors and that adaptive testing could substantially reduce test compromise, Marine Corps Headquarters tasked NPRDC with evaluating the feasibility of CAT for testing Marine recruits. The purpose of this paper is to report the results of the first in a series of studies investigating both the feasibility and the utility of CAT in comparison with a conventional test design.

The study was designed in part to address three research questions: (1) Is computer-based testing of military recruits administratively feasible? (2) Is a computer-administered adaptive test more reliable than a conventional test, holding test length constant? (3) If so, what is an appropriate length criterion for an adaptive test?

These questions were motivated by the results of previous research done elsewhere. The first question--that of administrative feasibility--seems trivial but is not. Interviews with military testing personnel indicated some misgivings about the ability of military recruits to use relatively sophisticated automated testing equipment, such as CRT computer terminals. This potential man-machine interface problem is the analogue of administrative difficulties encountered years earlier with paper-and-pencil tailored tests. For example, Seeley, Morton, and Anderson (1962) found that a substantial proportion of their military examinees did not successfully follow instructions on an experimental sequential item test; this experience may have caused a five-year lapse in military research on tailored or adaptive testing. Olivier (1974) had a similar experience using a paper-and-pencil flexilevel test in a sample of high school students.

The question of the advantages of adaptive tests over conventional ones in terms of reliability has a clear and positive theoretical answer: Holding test length and all else constant, a good tailored test design is superior, provided that highly discriminating test items are available (Urry, 1970).

This theoretical advantage is not always corroborated in empirical investigations. For instance, Bryson (1971) questioned the advantage of tailored testing over certain methods of conventional test design; Olivier (1974) failed to find an advantage for the flexilevel tests he used; and the results reported by Weiss and his colleagues have been less than unanimous in favor of adaptive tests. All these results are in contrast with those of Urry (1977), who reported that for his sample of 57 Civil Service job applicants an adaptive verbal ability test achieved an 80% reduction (compared to a conventional test) in the test length required to attain any of several specified levels of reliability. Urry's result was extraordinary. The only cloud over it is that it was based on indirect evidence: The conventional test reliabilities were based on Spearman-Brown equation adjustments to the reliability obtained in an independent sample, and the tailored test reliability was merely assumed, not rigorously verified.

Previous research into the reliability, validity, and efficiency of adaptive tests has often been inconclusive because of design flaws or nuisance factors. The major problem has been the lack of suitable means for estimating the adaptive test's reliability without making dubious assumptions. Another problem has been the general failure to match adaptive and counterpart conventional tests in item quality, with an unfair advantage usually in favor of the adaptive test. The research reported here was intentionally designed to remove those two problems--to provide credible indices of reliability that are appropriate for both test types and to provide a fair comparison by matching item quality across the test types. With those two problem sources eliminated, there is hope for an unequivocal comparison between adaptive and conventional test designs.

Method

The general method used was that of equivalent tests administered to independent examinee groups. One group took two equivalent computer-administered adaptive tests. The other group took two equivalent conventional tests, also administered by computer. In order to control for item quality, both test types were made up of items from the same source--a common pool of 150 verbal ability items, which had previously been calibrated in large samples of Marine recruits, using ICC methods.

Research Design

Each examinee was randomly assigned to one of the two treatment groups--Group A or C. Group A took two 30-item adaptive verbal ability tests, followed by a 50-item criterion test of word knowledge. Group C took two 30-item conventional verbal ability tests, followed by the same criterion test. All tests were administered at a computer terminal. Figure 1 is a schematic representation of the research design.

Observations. For each examinee who completed the tests, the following data were observed and automatically recorded:

1. Elapsed time for the testing session;
2. Elapsed time to complete pretest instructions;
3. Number of errors made during the instructions;
4. Number of times the proctor was called;

Figure 1
The Research Design for Administration
of the Experimental and Criterion Tests

Treatment Group	Tests: Adaptive Form 1	Tests: Adaptive Form 2	Tests: Conventional Form 1	Tests: Conventional Form 2	Tests: Criterion
A	X	X			X
C			X	X	X

5. Raw item scores (correct/incorrect);
6. Cumulative raw score after each item;
7. Latent trait ability estimates (experimental tests only);
8. Bayes posterior variance of the ability estimate after each item; and
9. Criterion test raw score.

The format for these observations is schematized in Figure 2.

Figure 2
Example of Examinee Record (Abbreviated)

	Raw Score		Ability Estimate		Posterior Variance	
Form:	1	2	1	2	1	2
Stage						
1	0	0	-.69	-.73	.548	.533
2	1	1	-.36	-.37	.401	.394
3	2	2	-.10	-.20	.332	.318
4	2	3	-.30	.02	.248	.266
5	3	4	-.14	.25	.229	.213
6	4	5	.01	.48	.193	.210
7	4	6	-.17	.65	.160	.184
8	5	6	-.05	.45	.145	.143
9	5	6	-.22	.26	.124	.115
10	6	7	-.15	.33	.115	.107
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
30	20	21	.59	.97	.053	.048

Criterion score	27
Total time	57.3 minutes
Instruction time	8.5 minutes
Instruction errors	1
Proctor calls	0

Independent variables. For the comparisons between the adaptive and conventional testing methods there were two independent variables: (1) test type (adaptive versus conventional) and (2) test length (5, 10, 15, 20, 25, 30 items).

Within the adaptive testing method, the test termination rule was treated as an independent variable for some analyses: Tests were terminated (1) at a fixed test length (5, 10, ..., 30 items) or (2) at a specified posterior variance (variable length). The number-of-items termination rule resulted, of course, in a test of predetermined length; and the posterior variance rule resulted in a variable length test, depending on the number of items required to attain specified levels of the Bayes posterior variance.

Dependent variables. Measures of the dependent variables were formed from the individual observations. The dependent variables included:

1. Testing time;
2. Instruction time;
3. Number of keyboard errors;
4. Number of proctor calls;
5. Alternate tests reliability coefficient after 5, 10, ..., 30 items; and
6. Test-criterion correlation after 5, 10, ..., 30 items.

Procedure

Items. The 150 items in the pool were calibrated using Urry's ancillary estimation method and were selected according to the prescriptions given by Urry (1977): All ICC slope parameters exceeded .80. The average value of the discrimination ($\underline{a}$) parameter was 1.24; item difficulty (location, or $\underline{b}$) parameters ranged from -2.0 to +2.0; and there were no items with a pseudo-guessing ($\underline{c}$) parameter greater than .30.

Examinees. Male Marine recruits reporting for duty at the Marine Corps Recruit Depot, San Diego, were the examinees. They were tested one at a time at a Burroughs TD832 terminal controlled by a Burroughs B1717 time-sharing minicomputer system. Assignment to groups (Group A or C) was randomized. Two hundred one examinees completed the tests--96 of these took the adaptive tests and 105 took conventional tests.

Tests. The conventional tests administered to Group C were rectangular tests spanning the difficulty range of the item pool. This broad range of difficulty was chosen in order to simulate the psychometric design of the verbal tests used in the ASVAB. Two 30-item equivalent forms--Form 1 and Form 2--were constructed from the 150-item pool. Items were chosen to be as highly discriminating as possible, consistent with the broad difficulty range. The two forms were constructed to be "weakly parallel" (Samejima, 1977), i.e., to have approximately equal test information functions. Within each form, the 30 items were sorted into five difficulty levels, then arranged in descending order of discriminating power within each level. The first five items in each form were the most discriminating items at their respective difficulty levels; items 6 through

10 were the second most discriminating items at each level; and so on. This arrangement resulted in two 30-item tests consisting of a sequence of six 5-item subsets each. This design was intended to permit meaningful analysis of the psychometric properties of rectangular conventional tests of lengths of 5, 10, 15, 20, 25, and 30 items. In order to equalize any effects due to test length, fatigue, or other extraneous factors, the two conventional tests were administered in counterbalanced item order, i.e., the two 30-item tests were administered as one 60-item test in the following order:

Item sequence:	1	2	3	4	5	6	7	8	...
Test Form:	1	2	2	1	2	1	1	2	...

The two 30-item adaptive tests were based on Owen's (1969, 1975) Bayesian sequential tailored testing procedure. For each examinee and each test form an initial normal prior distribution of ability was assumed, with mean 0 and variance 1.0. The test form (either 1 or 2) was counterbalanced for each examinee in a manner identical to that of the conventional tests: 12212112.... Both forms of the Bayesian test--Form 1 and Form 2--drew items from the same 150-item pool; counterbalancing the order of administration here served the added purpose of equalizing item quality across the two forms. The two adaptive tests were independent of each other except for their use of a common item pool.

The criterion test was formed by concatenating two obsolete operational test forms measuring word knowledge. This resulted in a 50-item test expected to be a highly reliable and fairly broad-range test of an important facet of verbal ability.

Results and Discussion

Feasibility

Data pertaining to the feasibility of using computer terminals to administer tests to military recruits are summarized in Table 1. Mean testing time was 61.0 minutes for the adaptive test group versus 50.4 minutes for the conventional test group. These were the mean times to answer 110 items--60 items from either the adaptive or the conventional alternate forms, followed by 50 criterion test items common to both groups. The adaptive tests required about 11 more seconds per item, or as much as 39% longer to answer than the conventional tests. Some or all of this difference may have been due to computations required for adaptive item selection, but this result does agree generally with Waters' (1977) finding that an adaptive test required significantly longer examinee processing per item than a similarly administered conventional test. In the present study, however, the observed time difference may be due in large part to idiosyncrasies of the computer system; if so, differences of the size reported here would not be expected if a faster computer were used to control and to administer the adaptive tests.

Instruction time averaged 9.5 minutes for the adaptive test group and 10.3 minutes for the conventional group; overall, the instructions required an average of 9.9 minutes. During this time, the examinees were familiarized with the CRT and keyboard by means of a programmed instructional sequence with special

Table 1
Testing Time and Examinee Error Summary for
Computer-Administered Test Sessions

	Group and Test		
Data	A Adaptive	C Conventional	Overall
Number of examinees	96	105	201
Mean time (minutes)			
Total	70.5	60.7	
Instructions	9.5	10.3	9.9
Testing	61.0	50.4	
Errors			
Procedural errors	25	30	55
Proctor calls	5	12	17

Note. Each session consisted of programmed instruction, 60 experimental test items, and a 50-item criterion test.

branching following procedural errors and with an audible call to the proctor if the examinee had difficulty correcting an error. Errors and proctor calls were counted. As the table indicates, there were 55 errors in all, in 201 test sessions; in only 17 cases was the proctor called. This amounts to about one procedural error per 4 test sessions and to a requirement for proctor intervention about one time per 12 test sessions.

Psychometric Characteristics

Reliability. Table 2 summarizes reliability and criterion validity data for both the adaptive and conventional alternate forms tests at lengths of 5, 10, 15, 20, 25, and 30 items.

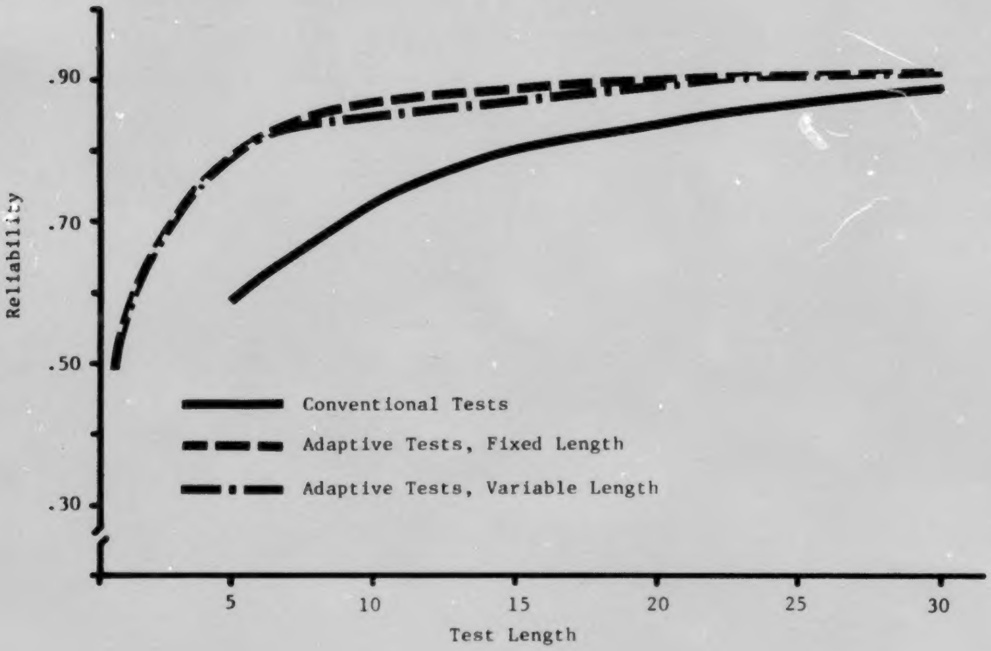
Table 2
Psychometric Characteristics of the Computer-Administered
Verbal Ability Tests as a Function of Test Type and Test Length

Psychometric Characteristic and Test	N	Test Length 5	10	15	20	25	30
Reliability							
Adaptive	96	.79	.87	.88	.90	.91	.91
Conventional	105	.59	.73	.80	.83	.86	.89
Validity							
Adaptive	93	.77	.82	.83	.84	.85	.85
Conventional	103	.73	.81	.84	.85	.85	.87
Relative efficiency		2.70	2.50	1.90	1.80	1.70	1.30

Reliability was operationalized as the correlation between scores on alternate forms at a given test length. The scoring procedure used was the same for both test types--latent ability estimation using the sequential estimation formulae developed by Owen (1969). From the table it is clear that the adaptive tests had substantially higher reliability coefficients than the conventional tests for any given test length. Viewing these data another way, it can be seen that the adaptive test reliability at a 5-item test length was practically equivalent to the conventional test's reliability at 15 items; similarly, the adaptive test's reliability at a length of 10 was superior to that of the conventional test at a length of 25.

Figure 3 contains a graphic comparison of the adaptive and conventional tests in terms of alternate forms reliability as a function of test length. Analysis of Table 2 and Figure 3 indicates that in terms of test length required to attain a given level of reliability, the adaptive tests had a substantial advantage over the conventional tests. This advantage was essentially the same for both fixed length and variable length stopping rules; there was no apparent advantage to variable length, as opposed to fixed length, within the adaptive testing method.

Figure 3
Alternate Forms Reliability Plotted as a Function of Test Length for the Conventional and Adaptive Tests



Relative efficiency. Thus, the adaptive tests achieved specific levels of reliability more efficiently than the conventional tests. How much more efficiently is indicated in row 3 of the table, labeled "relative efficiency." These data, based on the Spearman-Brown equation, estimate for each test length how much the conventional tests would have to be lengthened to attain the reliability of the adaptive tests. For example, the adaptive test reliability at 5 items, .79, was estimated to be equivalent to that of a conventional test 2.70 times as long, or 13.5 items in length. Notice that the relative efficiency of these adaptive tests always exceeds unity but diminishes as test length increases. Thus, the adaptive tests are more advantageous, at least in terms of relative efficiency, at fairly short test lengths. At lengths of 10 or fewer items, these adaptive tests were at least 2.5 times as efficient as the conventional tests. At lengths of 15 and more, however, the advantage, although still appreciable, is not quite so striking.

Validity. The advantage of adaptive tests was not so clear when the validity of the two test types is compared. Validity was operationalized as the correlation between test scores and the examinee's raw score on the concurrently administered 50-item Word Knowledge test. From their superior reliability, it would be expected that the adaptive tests would also be superior in validity at any constant test length. As Table 2 indicates, the adaptive tests had higher validities at test lengths up to 10 items; at lengths of 15 and up, however, the conventional tests had slightly higher validity. None of the validity differences was statistically significant at the .05 level.

Conclusions

Based on the data reported above, several conclusions are offered with regard to the feasibility and psychometric merits of adaptive aptitude testing of Marine recruits.

1. Testing Marine recruits with CRT terminals is feasible from both practical and human engineering standpoints. Embedded programmed instructions can effectively teach the recruits the use of the testing terminals. The number of proctors or attendants required to supervise and to assist in the testing room appears to be acceptably small.

2. Striking psychometric efficiency was demonstrated for the adaptive tests of verbal ability used in this study. It appears that in military personnel testing applications, well-constructed short adaptive tests can achieve high levels of measurement reliability with less than half the number of items required using conventional testing procedures.

3. There is no apparent psychometric advantage to the intuitively appealing notion of variable-length adaptive tests, at least for the adaptive testing method used here.

4. Short fixed-length adaptive tests of about 10 items per examinee seem to be sufficiently reliable for personnel testing purposes. The adaptive tests achieved a minimally satisfactory reliability level (.80) in just 5 items; additional test lengths beyond 10 items did not yield psychometric returns proportional to the added administration time required.

REFERENCES

Bryson, R. A comparison of four methods of selecting items for computer-assisted testing (Technical Bulletin STB 72-8). San Diego, CA: Naval Personnel and Training Research Laboratory, December 1971.

Gorham, W. A. Opening remarks. In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Symposium presented at the 83rd annual convention of the American Psychological Association, Chicago, 1975. Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, September 1976. (NTIS No. PB 261 694)

Olivier, P. An evaluation of the self-scoring flexilevel testing model. Unpublished doctoral dissertation, Florida State University, 1974.

Owen, R. J. A Bayesian approach to tailored testing (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service, 1969.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Samejima, F. Weakly parallel tests in latent trait theory with some criticisms of classical test theory. Psychometrika, 1977, 42, 193-198.

Seeley, L. C., Morton, M. A., & Anderson, A. A. Exploratory study of a sequential item test (Technical Research Note 129). Washington, DC: U.S. Army Personnel Research Office, December 1962.

Urry, V. W. A monte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 1977, 14, 181-196.

Waters, B. K. An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1977, 1, 141-152.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (NTIS No. AD 757788)

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesta, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

Weiss, D. J. Computerized adaptive ability measurement. Naval Research Reviews, 1975, 28, 1-18.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.

ACKNOWLEDGMENTS

This paper was also presented at the 1979 Conference of the Military Testing Association, San Diego, October 1979. Any opinions expressed are those of the author, and not necessarily those of the Department of Defense or the Department of the Navy.

PARALLEL FORMS RELIABILITY AND MEASUREMENT ACCURACY COMPARISON OF ADAPTIVE AND CONVENTIONAL TESTING STRATEGIES

MARILYN F. JOHNSON AND DAVID J. WEISS
UNIVERSITY OF MINNESOTA

Prior research at the University of Minnesota has compared the parallel forms reliabilities of adaptive and conventional vocabulary tests as a function of test length. The results are shown in Figure 1, which displays alternate forms reliabilities of Owen's Bayesian adaptive test and a conventional test as a function of number of items administered. The conventional test was peaked in information at $\theta = 0.0$; and test items were administered in order of information, from high to low values. The Bayesian adaptive test was scored by Bayesian methods; whereas the conventional test was scored by both proportion-correct and Bayesian methods. Both tests consisted of five-alternative multiple-choice vocabulary items.

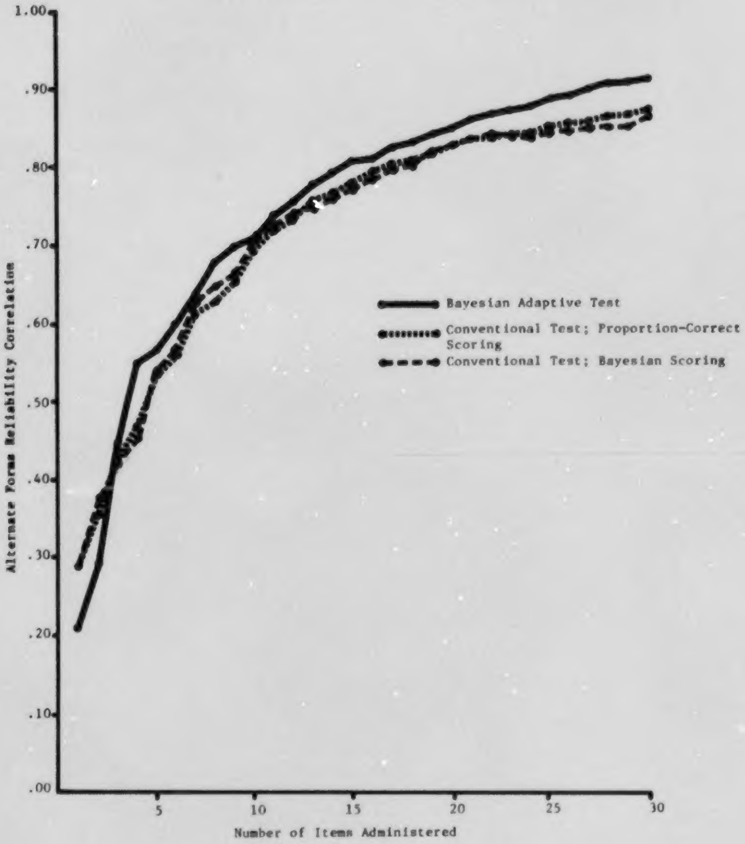
As expected, the plots in Figure 1 show an increase in reliability as test length increased for both testing strategies. However, rather than the expected asymptote of reliabilities for both strategies as test length increased, the reliability of the Bayesian adaptive test surpassed that of the conventional test. The approximate difference in reliabilities at test termination was $\underline{r} = .05$, with a 30-item reliability of .92 for the Bayesian test and .87 for the conventional test scored by the Bayesian method. The difference in reliabilities between Bayesian and proportion-correct-scored conventional tests was .04 at the 30-item test length.

The analysis also included a comparison of concurrent validity obtained by correlating the ability estimates with number-correct scores on a 120-item vocabulary criterion test also composed of five-alternative multiple-choice questions. These results (see Figure 2) indicated that although the Bayesian adaptive test was more reliable than the conventional tests, the conventional tests yielded higher validities when correlated with the criterion test. Figure 2 shows that the validities, similar to the reliabilities, increased as a function of test length, with the conventional test yielding higher validities after four items. The validity of the Bayesian test at 30 items was .797; that of the Bayesian-scored conventional test was .834; and the proportion-correct-scored conventional test obtained a validity of .841.

Purpose

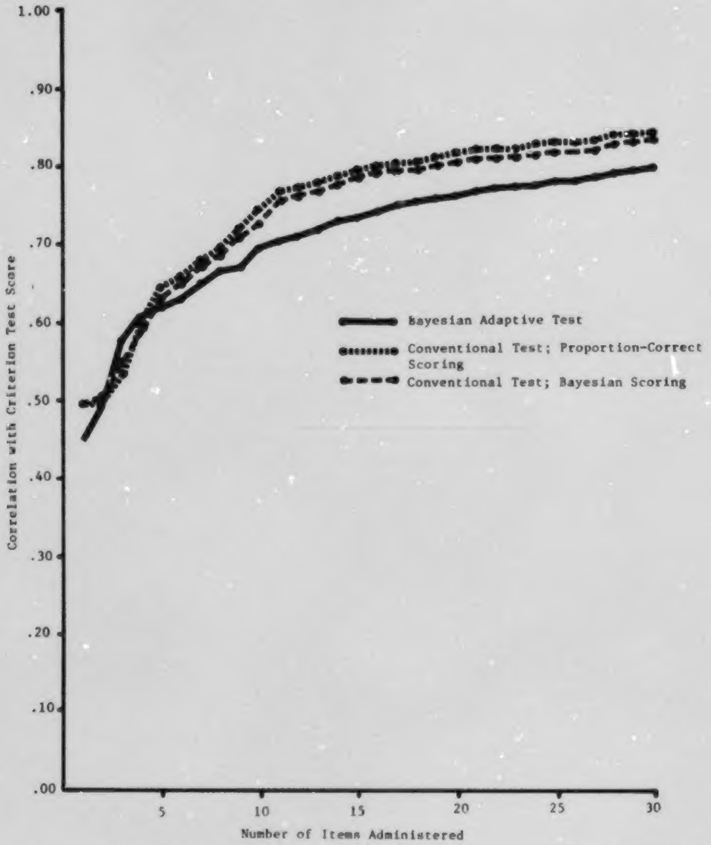
Due to the apparently contradictory nature of these findings, the present research was designed to replicate them. There were, however, some modifica-

Figure 1
Alternate Forms Reliabilities of Ability Level Estimates from a Bayesian Adaptive Test and a Conventional Test Scored by Proportion-Correct and Bayesian Scoring, as a Function of the Number of Items Administered



tions to the basic design of the comparison study, and an additional dependent variable, measurement accuracy, was used to compare the testing strategies. In addition, the present study compared peaked conventional, Bayesian adaptive, and maximum information adaptive testing strategies. The conventional test was also peaked in information evaluated at $\theta = 0.0$. Items on the conventional test were administered in order of item information but, for purposes of analysis, were arranged in random order. The item pool was composed of the same items that were used in the original study, but they were reparameterized after the original study and prior to the present investigation (Prestwood & Weiss, 1977). Comparisons of the three testing strategies were made in terms of parallel forms reliability as a function of test length and in terms of measurement accuracy as a function of θ level. Accuracy of measurement was operationalized as the posterior variance of the Bayesian-scored testing strategies and as standard errors of measurement for the maximum likelihood-scored testing strategies. Compari-

Figure 2
Correlations of Ability Level Estimates
from a Bayesian Adaptive Test and a Conventional Test
Scored by Proportion-Correct and Bayesian Scoring
with Criterion Test Score,
as a Function of the Number of Items Administered
(Averaged Across Two Test Forms)



sons of scoring strategies, including Bayesian, maximum likelihood, and proportion-correct scoring, were made on the basis of parallel forms reliability.

Method

Subjects

Undergraduate and graduate students from the University of Minnesota volunteered to participate in the fall 1978 and winter 1979 quarters. These students were recruited from Introductory Biology 1-011, Introductory Psychology 1-001, and a measurement course, Psychology 5-862. Students from the introductory psychology and biology courses participated in the study in order to obtain experimental points, which counted toward their final grade. Volunteers from the measurement course, both graduate and undergraduate students, participated at the request of the instructor.

There were 373 students in the conventional testing condition, 390 in the Bayesian testing condition, and 233 in the maximum information testing condition. Testing spanned two quarters in order to obtain an adequate number of students; a total of 996 students were tested during this period. Although students were recruited from varying subject pools, no difference in population was suggested because the undergraduate students were all from the College of Liberal Arts. In addition, students were sequentially assigned to one of the three testing strategies. The introductory biology and psychology students also participated in other studies during their experimental hour. In the case of the biology students, the experimental tests for this study were administered after a biology test. The fall 1978 introductory psychology students participated solely in this experiment, whereas the winter 1979 introductory psychology students first took the experimental test for this study, and then took another test. In each case, only data from the alternate forms verbal ability tests were analyzed.

Procedure

All students took the tests at an individual cathode-ray terminal (CRT) connected to a Hewlett-Packard real-time computer system. A test proctor was present during testing to provide assistance to the examinees. The students were assured that they could take as much time as necessary to complete the tests. Prior to administration of items on the first test, however, instructional screens explaining the operation of the CRTs were displayed. After students reviewed the test instructions and responded to a number of identification and demographic questions, the experimental tests were administered. Students responded to the five-alternative multiple-choice vocabulary questions by typing a number into the CRT corresponding to the chosen alternative.

Item Pools

Adaptive test. The Bayesian and maximum information tests used the same item pool from which to select items. The pool was composed of 256 items selected for the purposes of this study from the total vocabulary pool, which contained 358 items. The 358 items were newly parameterized items, based on combined data sources from conventional tests administered between fall 1969 and winter 1976. The items were parameterized with Urry's (1977) ESTEM program using a 3-parameter logistic ICC model. All items were assumed to have a guessing parameter of $\underline{c} = .20$. (Details regarding the parameterization procedure can be found in Prestwood & Weiss, 1977.) Selection of items from the larger pool was based on several criteria, which varied by difficulty levels of the items. Because there were few very difficult or very easy items, fewer items at these extremes on the difficulty continuum were eliminated. Items with discrimination parameters of $\underline{a} = 3.00$ were routinely rejected because this value was identified as a statistical artifact of the parameterization program and not as a true reflection of the item's discrimination value.

Based on a stratification of the items into difficulty levels, items were eliminated if their discriminations were low. This criterion, however, varied by difficulty level. In Levels 6 and 7, items were omitted if the discrimination parameter fell below $\underline{a} = .30$. In Levels 3, 4, and 5, where there were more items, the culling criterion was set at $\underline{a} = .35$. In these levels, also, items were omitted if the sample size on which the parameters were calibrated was less than 100. In many cases the items rejected on the basis of sample size were also of low discrimination.

Conventional test. The alternate forms of the conventional test were each composed of 30 vocabulary items arranged in descending order of item information evaluated at $\theta = 0.0$. The 60 most informative items at $\theta = 0.0$ were selected from the vocabulary pool composed of 256 items. By this procedure, items with relatively higher discrimination levels and difficulties of about $\underline{b} = 0.0$ were selected. Each test was thus peaked with respect to item information. Items were ordered by information at $\theta = 0.0$, and the 60 items were divided into Test Form A and Test Form B according to an ABBABAAB selection scheme. This procedure was used to insure that the alternate forms did not systematically differ in item information. The items were administered in order of descending item information. However, for purposes of analysis, pairs of items from the two test forms were randomly formed to simulate conventional paper-and-pencil testing conditions. The conventional test items were selected from the adaptive test pool so that it was possible that adaptive test items could also be used in the conventional test, since an independent groups design was being used.

Adaptive Testing and Scoring Strategies

Alternate forms of the adaptive tests were dynamically selected from the item pool by a special algorithm. Using an ABBABAAB rotational scheme, Form A of the adaptive test was given an opportunity to select an item from the pool of unadministered items, based on the item selection algorithm (Bayesian, maximum information) in use; and the ability estimate for that form of the adaptive test was updated. For administration of the next item to a testee, Form B then selected an item from the current pool of unadministered items; and the ability estimate for that form was updated. This procedure continued, using the ABBA-BAAB rotation, until 30 items were administered for each of the alternate forms--Form A and Form B--and the ability estimates for each form were saved after each item was administered.

Bayesian adaptive testing strategy. Items were selected and scored during the adaptive procedure according to Owen's (1975) Bayesian model. The prior distribution of ability was assumed to be normal, with a mean of 0.0 and a variance of 1.00. These values served as initial estimates of ability at the start of testing for each of the two forms for each individual. Testing was terminated after 30 items had been administered for each of the two forms. (Details concerning the Bayesian scoring algorithm can be found in McBride & Weiss, 1976.)

Maximum information adaptive testing strategy. Items were selected according to a maximum information item selection routine, and ability estimates were updated by scoring the responses by maximum likelihood methods (Bejar & Weiss, 1979). The initial estimate of ability was 0.0 for each form. Testing was terminated after 30 items had been adminstered for each of the two alternate forms.

The adaptive tests were scored after testing by a scoring strategy other than the one used during testing. The Bayesian test protocols were scored by maximum likelihood methods, and the maximum information test protocols were scored by Bayesian methods. Scores were calculated after each of the 30 items in both parallel tests. Responses to the two alternate forms of the conventional test were also rescored by Bayesian and maximum likelihood scoring methods at each test length from 1 to 30 items.

Independent Variables

Testing strategy was the major independent variable of interest. The strategies compared were the conventional, Bayesian, and maximum information testing strategies. Methods of scoring were also compared. These included logistic maximum likelihood scoring, Bayesian scoring, and (for the conventional test) proportion-correct scoring. Test length was a third independent variable of interest. Thirty test lengths were obtained by scoring each 30-item test 30 times. That is, a test was scored after the first item, after the first two items, after the first three items, and so on until 30 scores were obtained. In this way, 30 test lengths, varying from 1 to 30 items, were generated for each of the alternate forms.

Dependent Variables

Parallel forms reliabilities.

Testing strategies were compared on the basis of parallel forms reliability by correlating corresponding ability estimates obtained from Forms A and B for a given testing strategy. Since the test protocols were scored in at least two ways, Bayesian and maximum likelihood, a total of seven testing-scoring conditions were compared on the basis of parallel forms reliability. Scoring strategy was compared on the basis of parallel forms reliability by comparing reliabilities of a single testing strategy scored by more than one method. Three of the parallel forms reliabilities paired the appropriate scoring method with each of the three testing strategies. These were proportion-correct scoring of conventional tests, maximum likelihood scoring of maximum information tests, and Bayesian scoring of Bayesian-administered tests.

The remaining four parallel forms reliabilities were obtained by scoring the test protocols by a scoring routine other than the appropriate one. In this way, reliabilities were obtained for the Bayesian-scored maximum information test, the maximum-likelihood-scored Bayesian test, the Bayesian-scored conventional test, and the maximum-likelihood-scored conventional test. Proportion-correct scores were not obtained for adaptive tests. Reliabilities were calculated as a function of test length. That is, reliability was calculated not only from end-of-test ability estimates but also for each of the 30 test lengths. Scoring method correlations were obtained by correlating estimates obtained from different scorings of the same testing strategy. These correlations were used to analyze the similarity of ability estimates obtained from different scoring techniques applied to a single set of data.

Errors of measurement.

The three testing strategies were compared on the basis of their errors of measurement. This was assessed by two methods--one method estimated errors of measurement on the basis of maximum likelihood scoring methods; and the other, by Bayesian scoring methods. In the first method, test protocols were scored by maximum likelihood methods, and the standard errors of measurement (SEM) associated with each ability estimate was calculated. These values are the reciprocal of the square root of test information at a given θ level. They indicate how accurate the estimate is and how much it is likely to vary from the true θ value; the larger the standard error, the more likely the estimate will be inaccurate.

The SEM values were averaged within each of 20 $\hat{\theta}$ intervals ranging from approximately -3.0 to +2.0, and the mean SEM values were then plotted as a function of $\hat{\theta}$. This was done on a single randomly chosen parallel form for each of the three testing strategies.

The posterior variance of the Bayesian ability estimate was also used to compare the testing strategies on the basis of measurement accuracy. Posterior variances were averaged within each of 20 $\hat{\theta}$ intervals ranging from -2.0 to +2.0. These mean values were plotted at the midpoint of the $\hat{\theta}$ intervals and the points were connected to yield a continuous line. The posterior variance is analogous in meaning and interpretation to the standard errors of measurement.

Although one or the other of these measurement accuracy indices might have been adequate in comparing the testing strategies, both were included to minimize any biased conclusions regarding measurement accuracy of the adaptive tests. In general, posterior variance of Bayesian ability estimates will be less when items are selected according to a Bayesian testing strategy than when items are selected by any other adaptive procedure. Use of the posterior variance alone in the comparison of the adaptive testing strategies may bias conclusions toward the Bayesian testing strategy. For this reason the standard errors of measurement was also used as an index of measurement accuracy. This index, in general, will favor the maximum information testing strategy because items were selected and scored according to a maximum likelihood testing procedure.

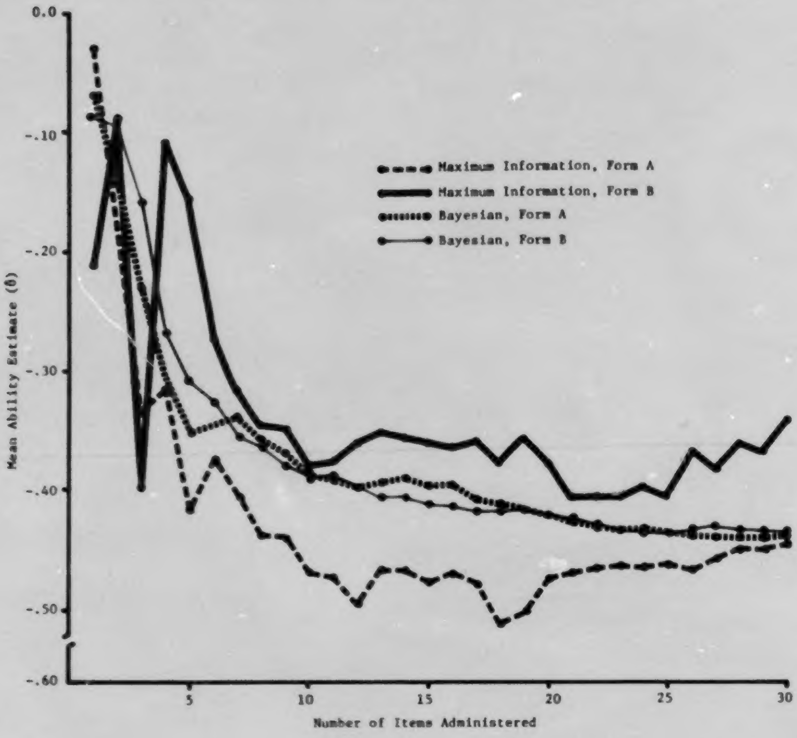
Results

Were the Tests Parallel?

Several analyses were performed to determine whether the alternate forms were functioning as parallel forms. These included comparisons of the means and variances of the ability estimates as a function of test length for the alternate forms of each testing strategy.

Score means. In general, the score means of the three testing strategies--conventional, Bayesian, and maximum information--showed an adequate level of parallel relationship between Forms A and B. Because the proportion correct score metric differs from the θ metric, the adaptive and conventional mean ability estimates are not directly comparable. Adaptive test comparisons of the means (Figure 3) show that there were greater differences between mean ability estimates for the alternate forms of the maximum information testing strategy than for the Bayesian testing strategy; this was because of the tendency of the Bayesian item selection and scoring routine to yield conservative estimates of ability. As testing progressed, however, differences between the ability estimates for the two alternate forms of each test decreased for both adaptive tests. Figure 3 also shows that the Bayesian mean ability estimates fell between the Form A and Form B means from the maximum information testing strategy. Thus, both adaptive procedures yielded about the same average ability estimates for the students selected from a common population.

Figure 3
Mean Ability Estimates from Parallel Forms A and B of Maximum Information and Bayesian Adaptive Tests, as a Function of Number of Items Administered

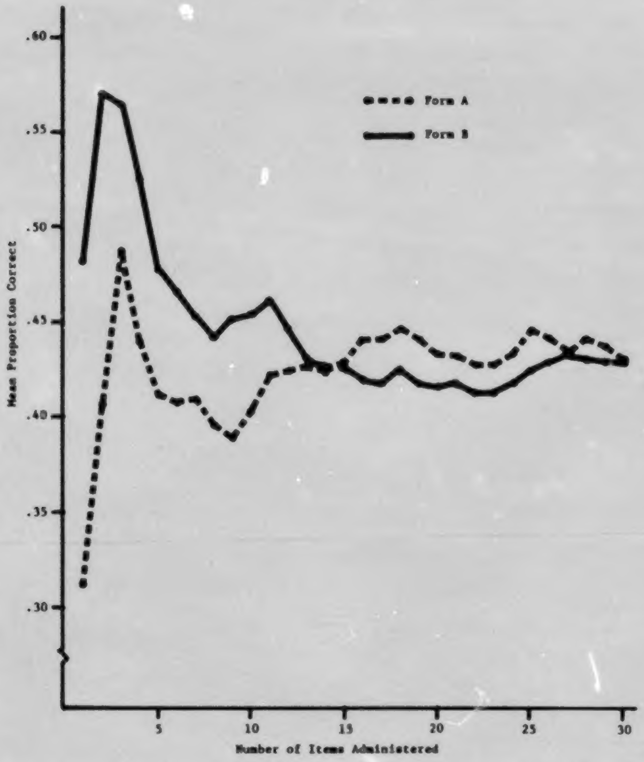


Means of the conventional parallel forms were obtained by averaging proportion-correct scores at each of 30 test lengths, based on randomly ordered items. Figure 4 shows that mean proportion-correct scores stabilized to a final value of .43.

Score variances. Variances of the ability estimates from the maximum information testing strategy (Figure 5) were relatively high up to 3 items, and then decreased steadily. The greatest difference in variance between the two alternate forms was at 3 items (1.25); whereas at 30 items the difference was only half (.75). Figure 5 also shows that ability score variances decreased from the beginning to the end of the test. Thus, score variances from the maximum information tests showed both a decrease in difference between alternate forms and a decrease in amount of variance as testing proceeded.

In comparison to the ability scores from the maximum information test, variance in Bayesian ability scores showed a similar maximum difference in vari-

Figure 4
Mean Proportion-Correct Score of the Conventional Test
for Alternate Forms A and B,
as a Function of Number of Items Administered

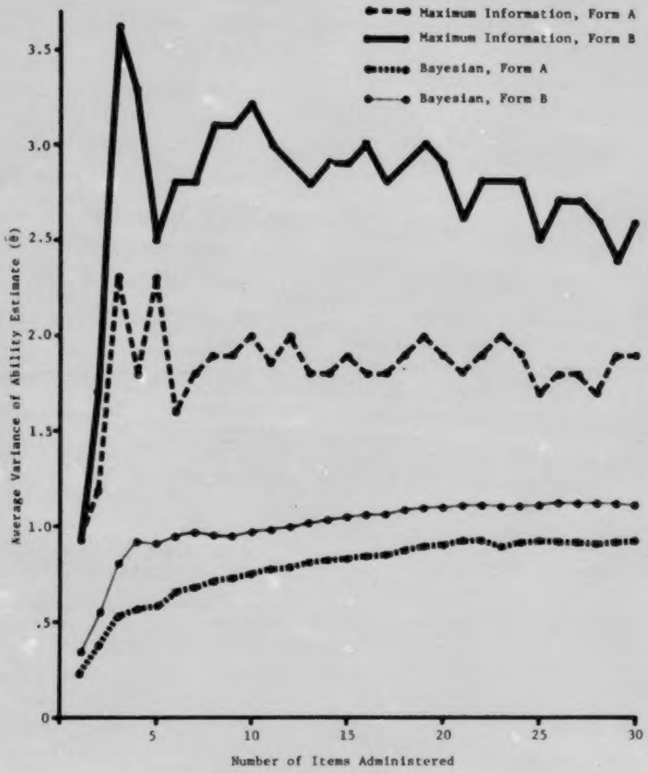


ance for tests of about 5 items in length, followed by decreased differences, as shown in Figure 5. Level of variance increased, however, as testing proceeded, reflecting the reduced dependence of the Bayesian ability estimates on the prior ability estimate. The restriction in Bayesian ability estimates due to the regression effect was still evident even at 30-item test lengths, since the ability estimate variances for the Bayesian tests were substantially lower than those of the maximum information tests.

Proportion correct score variance of both parallel forms of the conventional test decreased rapidly, from a possible maximum of .25 at 1 item to .06 at 30 items, as shown in Figure 6. Based on both the score means and score variances, the alternate forms of the conventional test were closer to being parallel than the alternate forms of either of the adaptive tests.

Errors of measurement as a function of test length. Samejima (1977) defines weakly parallel tests as tests that yield the same information functions. Thus, evidence for the parallel relationship between the adaptive forms included examination of their errors of measurement as a function of number of items administered. Average standard error of measurement, the reciprocal of the square root of theoretical test information, was used to compare alternate forms of the maximum information testing strategy. The error of measurement curves for the maximum information tests (Figure 7) showed the same form with variance decreasing rapidly to a final value of .40.

Figure 5
Average Variances of Ability Estimates for Forms A and B of Maximum Information Adaptive Tests and Bayesian Adaptive Tests, as a Function of Number of Items Administered

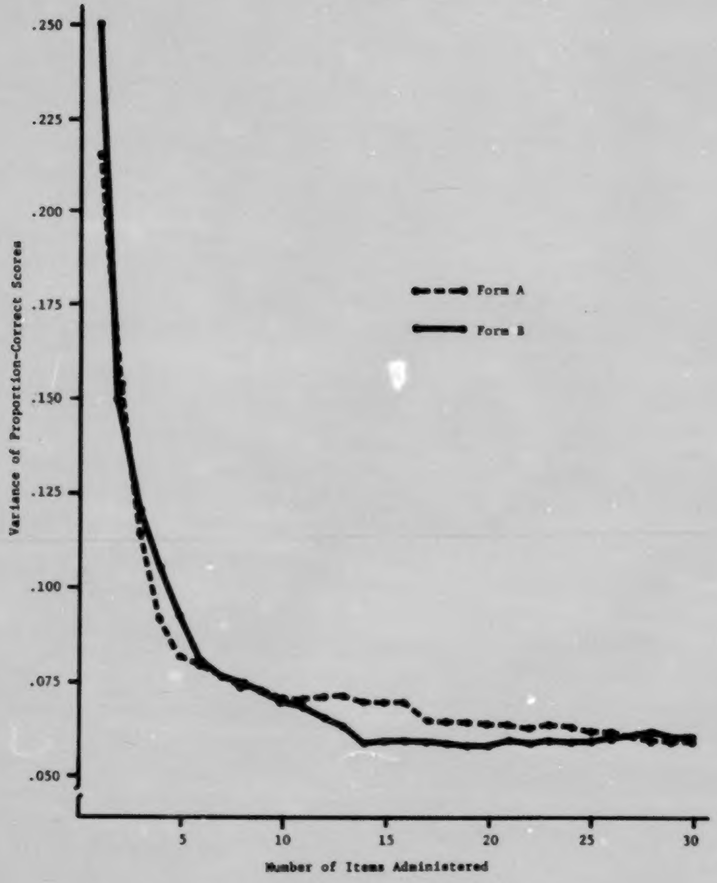


The error of measurement index for the Bayesian testing strategy was the posterior variance of the ability estimates. These data are also shown as a function of test length in Figure 7. Means of the Bayesian posterior variances for the two alternate forms were almost identical, decreasing from an initial value of .68, after 1 item was administered, to a final variance of .10, after 30 items were administered. As Figure 7 shows, there was less variance in Bayesian ability estimates than in the maximum likelihood ability estimates; but the data show that both the Bayesian and maximum information adaptive tests yielded parallel forms in terms of their mean errors of measurement, at almost all test lengths.

Parallel Forms Reliability

Optimal scoring method. The optimal scoring method was maximum likelihood

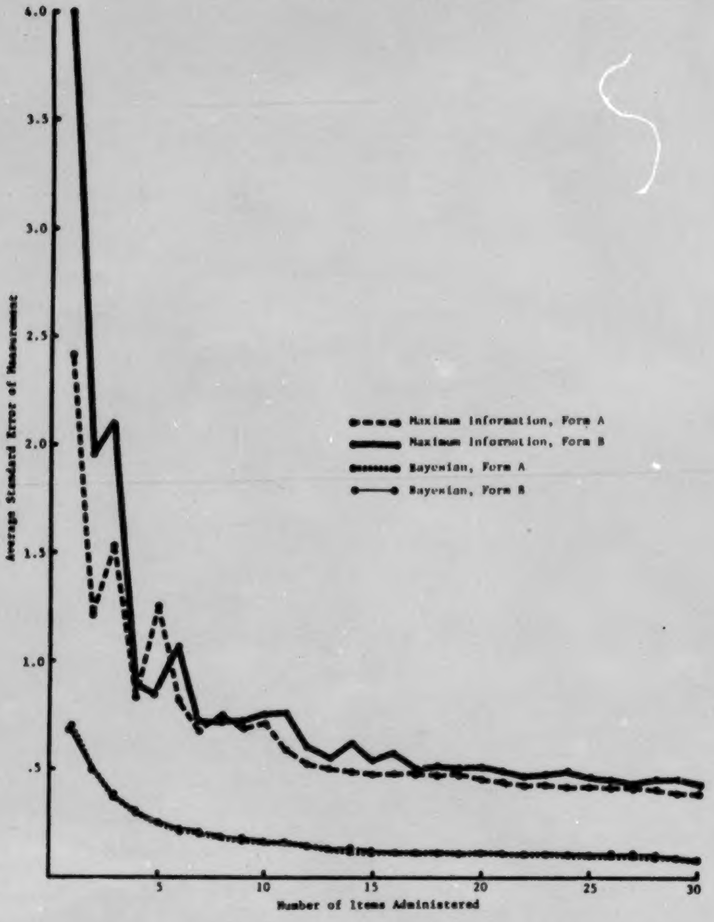
Figure 6
Variances of Proportion-Correct Scores from Alternate Forms A and B of the Conventional Test, as a Function of Number of Items Administered



for the maximum information testing strategy, Bayesian for the Bayesian testing strategy, and proportion correct for the conventional test. Alternate forms reliability correlations were computed at each test length for each testing strategy using these optimal scores.

Reliabilities of the three testing strategies as a function of test length are shown in Figure 8. The peaked conventional test yielded substantially higher reliabilities after 11 items than either of the adaptive tests. The greatest difference between reliabilities was $\underline{r} = .09$ between the adaptive and conventional tests at the 30-item test length; the reliabilities of the adaptive tests were $\underline{r} = .81$, compared with the final reliability of $\underline{r} = .90$ for the conventional test. The data in Figure 8 show essentially the same level and shape in reliabilities for the adaptive tests, although there was greater fluctuation in reliabilities for the maximum information test. The conventional test reliability was nearly identical to that of the Bayesian test up to the 10-item test

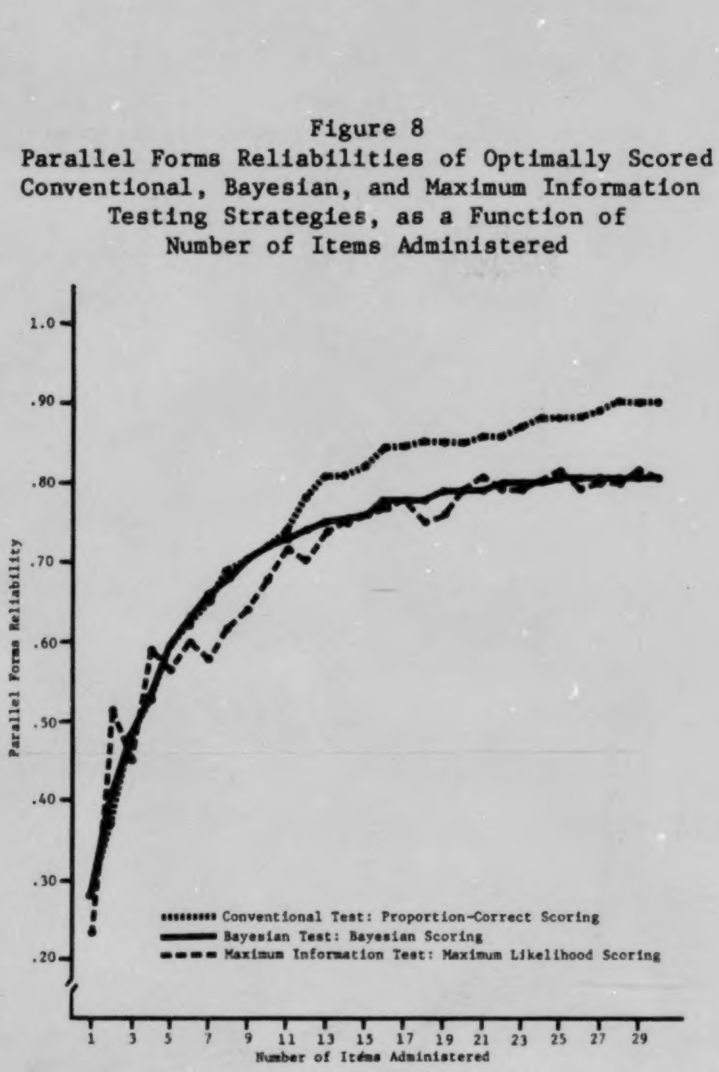
Figure 7
Means of Standard Error of Measurement from Parallel Forms A and B of Maximum Information Adaptive Tests and Mean Posterior Variance of Parallel Forms A and B of the Bayesian Adaptive Tests, as a Function of Number of Items Administered



length, but after that point the conventional test reliability increased more quickly than that of the adaptive tests. Although adaptive test reliabilities showed signs of leveling off toward the end of the test, the reliability of the conventional test seemed to increase steadily.

Other scoring strategy. Reliabilities were also obtained from testing strategies scored by other than optimal scoring strategies. Four testing-scoring combinations were of interest: Bayesian-scored maximum information tests, maximum-likelihood-scored Bayesian tests, Bayesian-scored conventional tests, and maximum-likelihood-scored conventional tests. These reliability results are shown in Figure 9 as a function of test length.

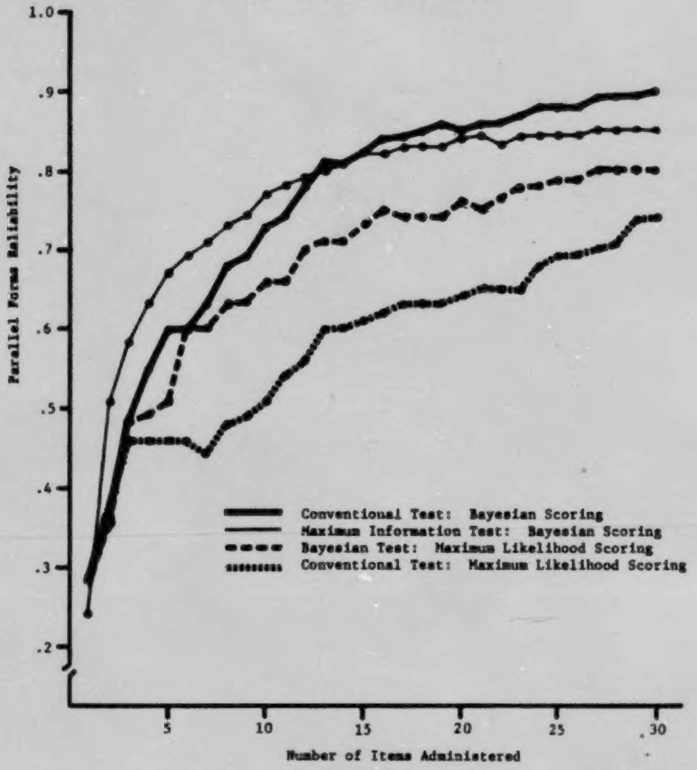
Figure 8
Parallel Forms Reliabilities of Optimally Scored Conventional, Bayesian, and Maximum Information Testing Strategies, as a Function of Number of Items Administered



In general, Figure 9 shows that the Bayesian scoring procedure yielded higher reliabilities under nonoptimal conditions than the maximum likelihood scoring procedure. Bayesian scoring of the conventional test yielded essentially equivalent reliabilities at every test length, as did proportion-correct scoring of the conventional test. Bayesian scoring of the maximum information tests yielded higher reliabilities at most test lengths beyond about 12 items than the optimal scoring strategy for that test. In addition, Bayesian scoring of the maximum information test tended to decrease substantially the differences in reliabilities observed between the conventional and adaptive tests. Figure 9 shows that the reliability for the Bayesian-scored maximum information test was higher than that of the conventional test for test lengths from 3 to 12 items. The maximum difference between these two reliabilities was $\underline{r} = .05$ at 30 items, as compared to $\underline{r} = .09$ for the data in Figure 8. These data indicate that Bayesian scoring of an adaptive test may yield more stable estimates of ability than maximum likelihood scoring.

The data also illustrate the inappropriateness of scoring conventional

Figure 9
Parallel Forms Reliabilities of Non-Optimally Scored Testing-Scoring Strategies, as a Function of Number of Items Administered



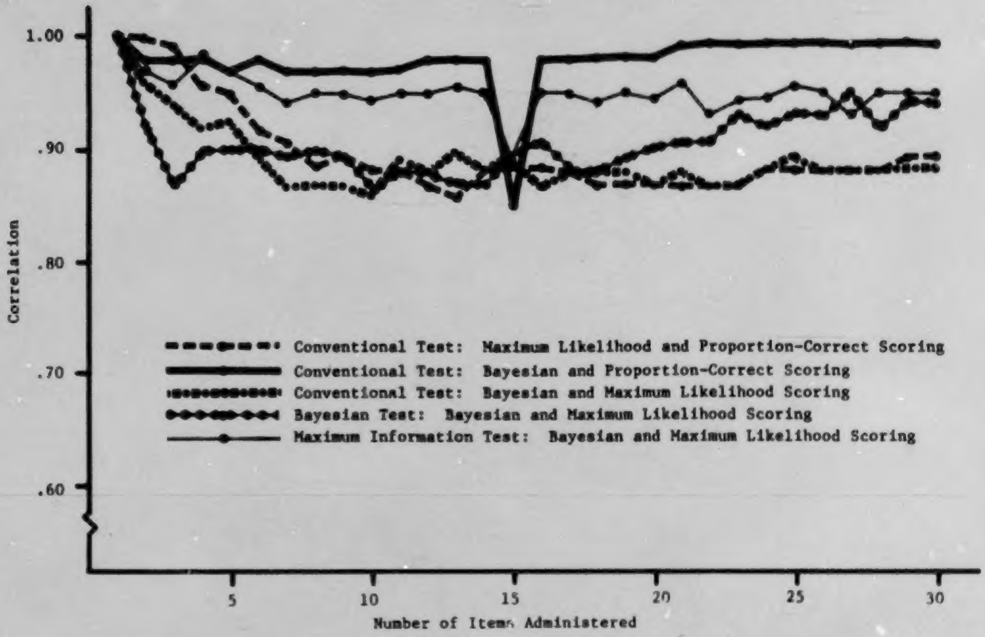
tests with maximum likelihood scoring methods. As Figure 9 shows, maximum likelihood scoring of the conventional test resulted in extremely low reliabilities at all test lengths, reaching a maximum of only .74 at 30 items.

Scoring Method Correlations

To study the generality of the findings of Kingsbury and Weiss (1979), in their study of correlations among latent-trait scoring methods in achievement test data, comparisons of the ability estimates from the various scoring methods were made by correlating scores obtained from different ways of scoring the same testing strategy. For both adaptive testing strategies, Bayesian scores were correlated with maximum likelihood scores. Conventional test comparisons were made by correlating proportion-correct scores with Bayesian scores, proportion-correct scores with maximum likelihood scores, and Bayesian scores with maximum likelihood scores. For each testing strategy, one of the two alternate forms was randomly chosen for these analyses. These five scoring combinations are shown in Figure 10 as a function of test length.

As Figure 10 shows, the highest correlations were between Bayesian and proportion-correct scores of the conventional test. These correlations varied in

Figure 10
Correlations Between Scoring Methods
for the Same Alternate Form, as a
Function of Number of Items Administered



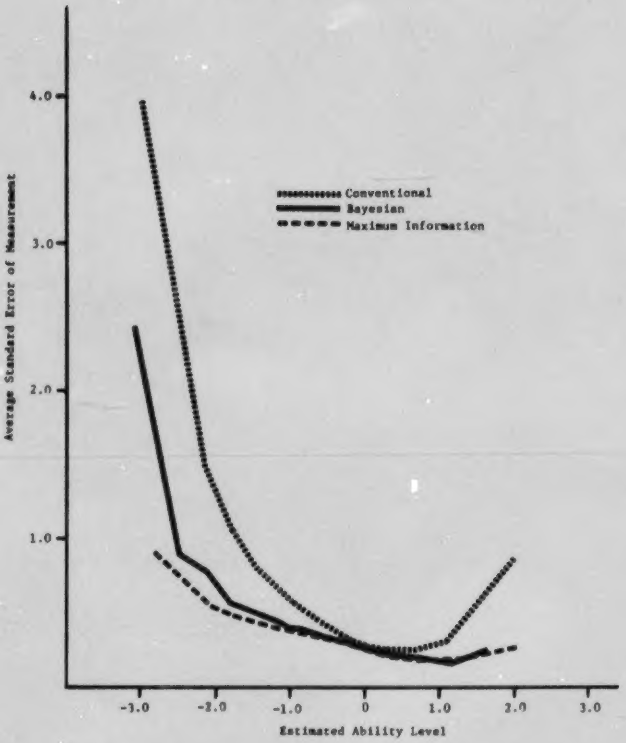
value between 1.00 for a 1-item test to .85 for a 15-item test, with most correlations between .97 to .99. The second highest level of correlation was between the Bayesian- and maximum-likelihood-scored maximum information test, with most correlations between .93 and .95. With the exception of the latter half of the correlations between Bayesian and maximum likelihood scores from the Bayesian test, there were few differences among the other three sets of correlations; the modal correlation for these three plots was .88. The correlations between Bayesian and maximum likelihood scores from the Bayesian test increased steadily after the 15-item test length to a final value of $\underline{r} = .94$.

Measurement Precision as a Function of Ability Level

Figure 11 shows plots of the average standard errors of measurement as a function of the maximum-likelihood-derived ability distribution. These data are the reciprocal of the square root of the test information function for each test. The distribution obtained from this sample varied from about -3.00 to +2.00 and was divided into equal frequency intervals ($N \geq 20$), separately for each testing strategy.

The data indicate that at no point on the ability continuum were the standard errors of measurement smaller in the conventional test than in the adaptive tests. In general, the maximum information testing strategy yielded smallest standard errors or greatest measurement precision. The Bayesian test, when scored by maximum likelihood, had poorer measurement precision at the lower ex-

Figure 11
Average Standard Error of Measurement as a Function of Ability Level for Conventional, Bayesian, and Maximum Information Testing Strategies
(Non-Converging Values Eliminated)

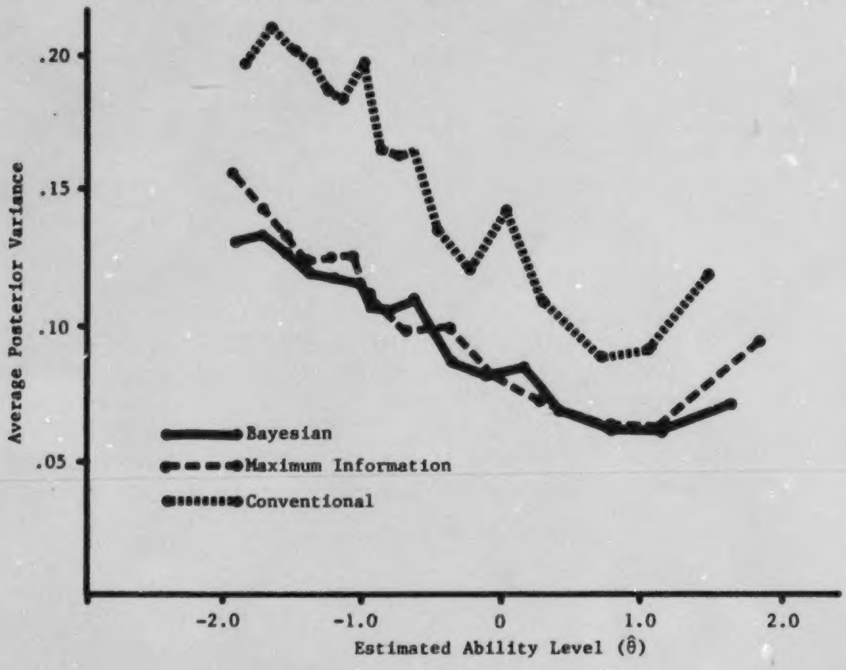


treme of the ability continuum than did the maximum information test. Precision of measurement for all the testing strategies was greatest at the central portion of the ability distribution than at the extremes.

Bayesian posterior variance comparisons are shown in Figure 12 as a function of the Bayesian-derived ability distribution. The distribution varied from about -2.00 to +2.00. The average posterior variance was greater at all points along the ability continuum for the conventional strategy than for either of the adaptive tests. The Bayesian and maximum information testing strategies had about the same level of measurement accuracy in the center of the ability distribution. At the extremes of the ability continuum, the Bayesian testing strategy resulted in slightly better measurement precision than did the maximum information testing strategy.

In both error of measurement comparisons, there was poorer measurement at the low end of the ability distribution, although the extremes--both positive and negative--were less precisely measured than the center of the ability continuum. The results indicate that the adaptive tests yield about the same level of measurement precision and that these levels were greater than those obtained from the conventional test at all levels of ability.

Figure 12
Average Bayesian Posterior Variance of
Ability Estimates as a Function of
Ability Level for Conventional, Bayesian,
and Maximum Information Testing Strategies



Discussion

The major finding in this study was that the conventional test yielded higher alternate forms reliability than did the adaptive tests. However, when the maximum information adaptive test was scored by the Bayesian scoring algorithm, reliabilities of short adaptive tests were higher than those of the conventional test, and differences in reliabilities were smaller at longer test lengths. Limitations of the item pool might account in part for the lowered reliability of the adaptive tests in comparison to the conventional test, since adaptive tests depend heavily on the quality of the items in the item pool. When an item pool consists of highly discriminating items, every ability level along the latent trait continuum can be measured with a high degree of precision using adaptive tests (McBride & Weiss, 1976). When there are few items to measure abilities at the extremes and/or the available items are of low discrimination, abilities at the extremes cannot be measured accurately.

The item pool used for the two adaptive tests had fewer items at the extremes of the ability range and these items had relatively lower discrimination parameters. It is likely that, especially at abilities where there were fewer items, the correlations between ability estimates would be attenuated and the adaptive process would be at a disadvantage as testing progressed. The result would be that toward the end of testing there would be fewer and fewer items available at a given ability level.

The adaptive test scoring process also depends on accurate parameterization of items and on testees responding according to a single latent trait. Experimental subjects taking a test that does not relate to any course they are taking and that does not count for a grade may respond carelessly, with less than full attention. It is unknown to what extent the item parameters are inaccurate. An optimal research strategy for comparison of conventional, Bayesian, and maximum information testing strategies on the basis of parallel forms reliability is through simulated testing. The disadvantage of inaccurate item parameters, nonoptimal item pool characteristics, and the possibility that students did not respond exclusively in accordance with their ability level can be alleviated in simulation.

One additional factor that limits the comparison of the testing strategies in terms of alternate forms reliability correlations is the distribution of ability in the population. Since values of the Pearson product-moment correlations depend on the distributions of the ability estimates involved, different ability distributions can result in different levels of correlation. Thus, the reliability correlations confound the distribution of the ability estimates with the measurement precision of the testing strategies. Information is a measure of precision of measurement, yielding comparisons of testing strategies that are unconfounded by the distribution of the ability estimates. As Figure 11 shows, both adaptive testing strategies yielded scores with greater precision/information (lower errors of measurement) than did the conventional testing strategy.

On the basis of the reliability data, few conclusions can be drawn about the relative merits of the adaptive testing procedures. Bayesian scoring of the Bayesian test showed higher reliability than the maximum-likelihood-scored maximum information test. Bayesian scoring of the conventional and maximum information testing strategies yielded higher reliabilities than maximum likelihood scoring of the conventional and Bayesian testing strategies. This might indicate either that the Bayesian scoring algorithm yields more reliable estimates of ability or that it yields the same regressed or biased estimate of ability. The Bayesian test would tend to yield higher parallel forms reliabilities than the maximum information testing strategy in the case where most items measuring abilities at the extremes of the distribution are of lower discrimination. Because the Bayesian adaptive test yields regressed estimates of ability and requires fewer items measuring abilities at extreme θ values, the Bayesian ability estimates obtained, although biased, would be more stable than ability estimates from the maximum information testing strategy.

REFERENCES

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1980. (NTIS No. AD A067752)

Kingsbury, G. G., & Weiss, D. J. A comparison of a Bayesian adaptive testing strategy to a conventional testing strategy: Alternate-forms reliability and criterion validity. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, in preparation.

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Prestwood, J. S., & Weiss, D. J. Accuracy of perceived test-item difficulties (Research Report 77-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, May 1977. (NTIS No. AD A041084)

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (Ed.), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, September 1976. (NTIS No. PB 261 694)

ACKNOWLEDGMENTS

This study was supported by funds from Army Research Institute, Air Force Human Resources Laboratory and Office of Naval Research, under Contract N00014-79-0324 NR 150-431 with the Personnel and Training Research Programs of the Office of Naval Research.

A COMPARISON OF THE ACCURACY OF BAYESIAN ADAPTIVE AND STATIC TESTS USING A CORRECTION FOR REGRESSION

STEVEN GORMAN
DEPARTMENT OF THE NAVY

The vast changes in computer technology have made a strong impact upon the field of ability measurement. The increased capabilities and decreased costs of computer use have opened the door to application of latent trait theory. Two Bayesian procedures for ability estimation have become popular--the Bayes modal procedure (Samejima, 1969) and the Owen (1975) algorithm. Both Bayesian procedures use a prespecified distribution, usually the Gaussian normal distribution, as the prior variance of ability. The item characteristic curve (ICC; also called the item response function) is employed as the likelihood function. The product of the prior distribution and the likelihood function is the posterior distribution of ability. These two procedures can be used in either conventional or adaptive mode.

McBride and Weiss (1976) have studied Owen's Bayesian adaptive procedure and have determined that with this procedure, ability estimates regress toward the mean. That is, high-ability examinees tend to achieve lower ability estimates, and low-ability examinees tend to have higher ability estimates. Urry (1977) has suggested a correction, namely, dividing the Bayesian regressed ability estimate by the test reliability. A second, potentially more serious, problem is the reliance upon accurate 3-parameter logistic item parameters. Urry (1976) developed OGIVIA3, a computer program to estimate these item parameters. The effectiveness of this estimation procedure for use in the Owen algorithm was reviewed by Gugel, Schmidt, and Urry (1976). OGIVIA3 has been revised (Croll & Urry, in prep.) and has been renamed ANCILLES.

The purpose of the present paper is to evaluate the effectiveness of two Bayesian ability estimation procedures with a correction for regression using known and estimated parameters. Specifically, the studies simulated the Owen algorithm and Bayes modal testing methods in both adaptive and static mode with a correction for regression using known parameters and the parameters estimated using ANCILLES.

Study 1: An Analysis of the Verbal Scholastic Aptitude Test

Background and Purpose

Lord (1968) applied the 3-parameter logistic model developed by Birnbaum (1968) to the Verbal Scholastic Aptitude Test (VSAT). Until Lord's article, little research had been conducted using Birnbaum's model. However, since this article, with the exception of a few articles involving the maximum likelihood procedure (Bejar, Weiss, & Gialluca, 1977; Kolakowski & Bock, 1970, 1972; Wood, Wingersky, & Lord, 1976), the overwhelming majority of latent trait research has applied the work of Birnbaum to adaptive tests and not to conventional tests. Samejima (1968) detailed the mechanics of a Bayes ability estimator based on a response pattern of test items. She proved that with an assumed normal distribution of ability as a prior distribution, and using the ICC as a likelihood function, the mode of the posterior distribution will provide an absolute maximum, which can be used as an ability estimate. Urry (1976) incorporated the Bayes modal procedure in the second stage of his item parameter estimation program. Owen (1975) developed a Bayesian procedure for estimating ability; however, this procedure was developed for the adaptive mode. Bejar and Weiss (1979) programmed the Owen algorithm for scoring static tests, but no data on its effectiveness were made available.

The purpose of this study was to investigate the efficiency of the Bayes modal and Owen's Bayesian ability estimation procedures relative to a conventional rights-only scoring. In particular, the issues investigated are (1) conditional bias, (2) conditional accuracy, and (3) precision of test scores.

Design of the Study

Artificial data were generated according to the 3-parameter logistic model:

$$P_i(\theta) = c_i + (1 - c_i)\ [1 + \exp(-1.7a_i(\theta - b_i))]^{-1} \qquad [1]$$

using the LVGEN program developed by Urry (1971). This program provided vectors of responses, correct (1) or incorrect (0), for the simulated examinees (sims). The test items used had the parameters of the first 80 VSAT items reported in Lord (1968).

For the purpose of this study, it was assumed that the item parameters reported in Lord's study were the actual parameters and not estimated, as they actually were. The 80 item parameters were administered to 2,000 sims from a normal distribution (mean 0, variance 1) generated by the LLRANDOM Computer Program (Learmonth & Lewis, 1973) in conjunction with the LVGEN program. The resulting vectors of simulated binary responses were analyzed by the ANCILLES Program; estimates of the 80 "known" VSAT parameters were the resultant output. This allowed a comparison of the robustness of the Bayesian ability estimation programs to inaccuracy in the item parameter estimates. An additional 2,000 normally distributed sims were administered the VSAT items. This permitted computation of the correlation of known ability with the various ability estimates and the mean and variance of raw scores so that a Z-transformation could be computed. This allowed comparison of a simpler scoring procedure based on classical test theory with the two scoring procedures based on latent trait theory.

Five conditions of scoring the same item responses were examined: (1) Bayes modal ability estimates based on known item parameters, (2) Bayes modal ability estimates based on estimated item parameters, (3) Owen's Bayesian ability estimates based on known item parameters, (4) Owen's Bayesian ability estimates based on estimated item parameters, and (5) ability estimates based on raw score to Z-score transformations.

To properly address the evaluation mentioned above required examination of the test score characteristics as a function of ability level. Therefore, the ability distribution consisted of 100 sims at each of 11 equally spaced values in the interval $-2.5 \leq \underline{b} \leq +2.5$.

For each of the five simulated test administrations, conditional bias, conditional accuracy, and conditional precision were estimated from the 100 observations at each ability level (θ_e).

Conditional bias. This statistic provided an indicator of the magnitude and direction of the error between true ability and ability estimated by each of the scoring procedures at various levels of the trait continuum where

$$\text{bias} = b_e | \theta_e = \bar{\hat{\theta}}_e - \theta_e \, , \qquad [2]$$

where

$\underline{b}_e$ = average bias for each of 11 values of on the trait continuum,
θ_e = true ability of examinees for each value, and
$\hat{\theta}_e$ = average ability estimates for each value.

Conditional accuracy. The accuracy of the test scores was provided by the root mean square error computed for the 11 values using the formula

$$e_i | \theta = \left[\sum_{i=1}^{n} (\theta_e - \hat{\theta}_e)^2 \, n^{-1} \right]^{\frac{1}{2}} \qquad [3]$$

where

$\underline{e}_i | \theta$ = root mean square error conditional upon ability level,
$\underline{n}$ = 100,
$\bar{\theta}$ = known ability level, and
$\hat{\theta}_e$ = the ability estimate.

Conditional precision. This statistic was provided by the test score information function. The information generated by a score about a given ability level can be compared to the precision of measurement at that point. Samejima (1977) stated that the inverse of the square of information can be considered as the standard error of measurement when number of items and test information are sufficiently large. Birnbaum (1968) provides a formula for information:

$$I_e(\theta') = \left[\frac{\partial}{\partial \theta} \, \frac{E(\hat{\theta}_e | \theta}{\hat{\sigma}_{\hat{\theta}_e} | \theta} \right]^2 \qquad [4]$$

where I_e (θ') is the information about θ provided by score $\underline{x}$. Sim scores were calculated at each of 11 equally spaced ability levels $-2.5 \leq \underline{x} \leq +2.5$; these test score means were used to estimate the slope by fitting a curve through three consecutive values. Because test score means were required on either side of the information point, information values could not be computed for the ends of the continuum (-2.5, +2.5).

Results

Estimation bias. The comparisons between the two Bayesian procedures for scoring static tests using estimated parameters and the raw score to Z-score transformation are in Figure 1. The figure shows that the absolute value of bias for the Z-score was much greater than for the two Bayesian procedures at ability level -2.5. The absolute value of Bayesian score bias tended to be equal to or lower than that of the Z-score along the entire trait continuum. Of the two Bayesian procedures, the Bayes modal bias was greater at upper trait levels.

Figure 1
Bias of Three Scoring Procedures, Using
Estimated Item Parameters in a Static Test

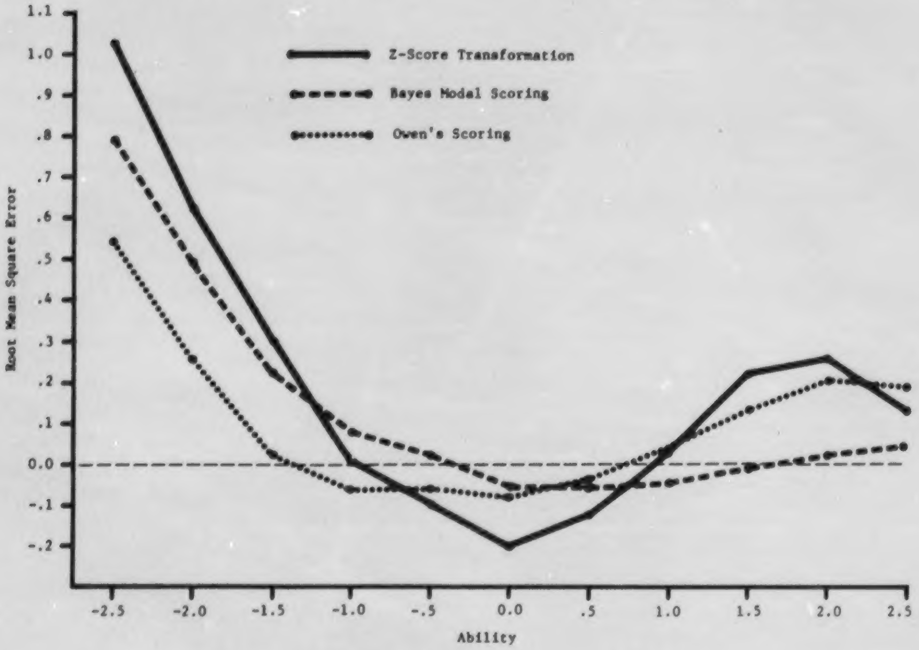
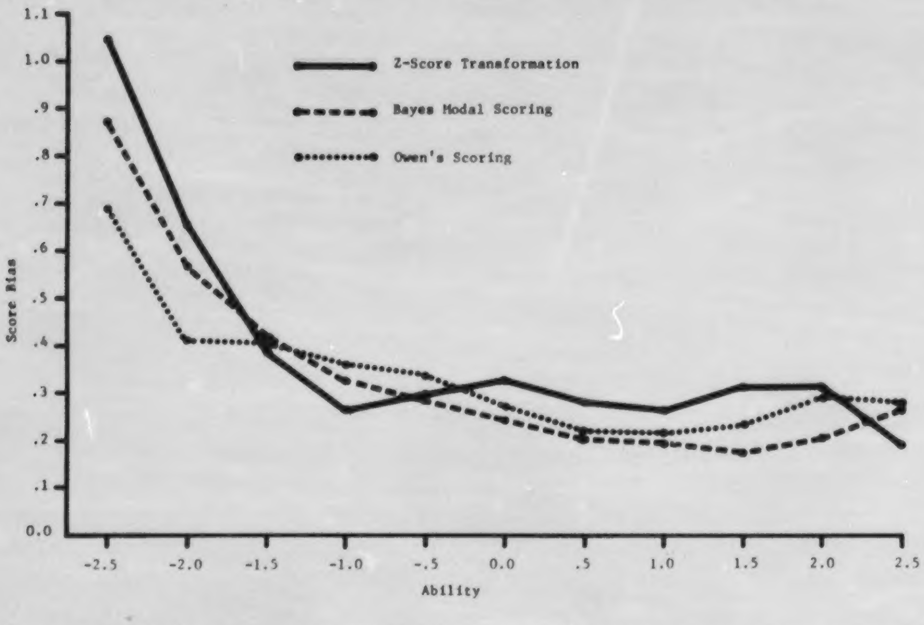


Table 1 shows the bias values of the two Bayesian procedures under conditions of known and estimated parameters, as well as the conventional Z-score method. The Bayes modal scores using known parameters still suffered to some degree from the regression to the mean effect, although deviations from zero were mostly lower than the bias from either estimated Bayes or Z-score methods. Improvements to the estimation of item parameters could decrease the bias of the two Bayesian static procedures significantly.

Table 1
Bias of Conventional Z-Score Method and Two Bayesian Scoring Methods Using Estimated and Known Parameters in a Static Test

Ability Level	Z-Score	Estimated Bayes Modal	Estimated Owen´s	Known Bayes Modal	Known Owen´s
-2.5	1.027	.789	.541	.482	.366
-2.0	.621	.488	.253	.260	.059
-1.5	.300	.221	.022	.087	-.143
-1.0	.010	.077	-.062	.065	-.142
-0.5	-.097	.022	-.058	.051	-.083
0.0	-.200	-.056	-.081	-.019	-.064
0.5	-.126	-.058	-.035	-.023	-.008
1.0	.031	-.048	.037	-.027	.046
1.5	.220	-.013	.132	-.030	.104
2.0	.260	.023	.206	-.044	.138
2.5	.130	.044	.188	-.060	.109

Figure 2
Root Mean Square Error of Three Scoring Methods Using Estimated Item Parameters in a Static Test



Conditional accuracy. Figure 2 displays the root mean square error (RMSE) of ability estimation for the two Bayesian algorithms using estimated parameters and the Z-score method. All three methods followed the same trend of having high RMSE values at the low-ability levels and diminishing asymptotically to a value of about .2 at the trait level +.5. This phenomenon appeared to be a function of the test itself, with its emphasis on more precise measurement at the higher ability levels. The conventional scoring procedure tended to have the highest inaccuracy, with two exceptions (ability levels -1.0 and +2.5). Table 2 lists the RMSE values for the two Bayesian methods using known and estimated parameters.

Table 2
Root Mean Square Error of the Z-Score Method
and Two Bayesian Scoring Methods Using
Estimated and Known Parameters in a Static Test

		Parameters			
		Estimated		Known	
Ability Level	Z-Score	Bayes Modal	Owen's	Bayes Modal	Owen's
-2.5	1.048	.875	.686	.703	.537
-2.0	.652	.567	.412	.486	.365
-1.5	.386	.418	.405	.482	.463
-1.0	.263	.325	.361	.370	.425
-0.5	.296	.284	.338	.300	.382
0.0	.328	.243	.273	.241	.288
0.5	.281	.203	.221	.191	.213
1.0	.264	.195	.215	.185	.212
1.5	.313	.176	.233	.160	.204
2.0	.314	.205	.293	.188	.239
2.5	.191	.266	.280	.252	.231

Conditional precision. The test score information values at the nine ability levels, -2.0 to +2.0, for the two Bayesian scoring methods using estimated parameters and the conventional scoring procedure, are in Figure 3; numerical values are in Table 3. The data in Table 3 coincide with two trends of the earlier study (Lord, 1968, p. 998) on the VSAT. First, the data in Table 3 (as well as in Table 2) illustrate the more precise measurement on the VSAT at upper ability levels. Second, the data show that significant increases in precision can be gained by using the Bayesian scoring procedures.

The original study weighted items based on the logistic model and found this procedure provided greater information than conventional scoring. The average score information value for conventional scoring was 12.195; the average for the Owen scoring was 13.800 and was 14.120 for the Bayes modal scoring, with estimated item parameters used in the scoring procedures. Slightly higher averages (13.960 for the Owen and 14.503 for the Bayes modal scoring) occurred when the known item parameters were available.

Fidelity. Fidelity coefficients, the correlations of the known ability of 2,000 sims from a normal population with their estimated abilities, were computed from the various test scoring methods and are in Table 4. Although the increase in the correlation is only roughly .02 for the two Bayesian methods over

Figure 3
Test Score Information of Three Scoring Methods, Using Estimated Item Parameters in a Static Test

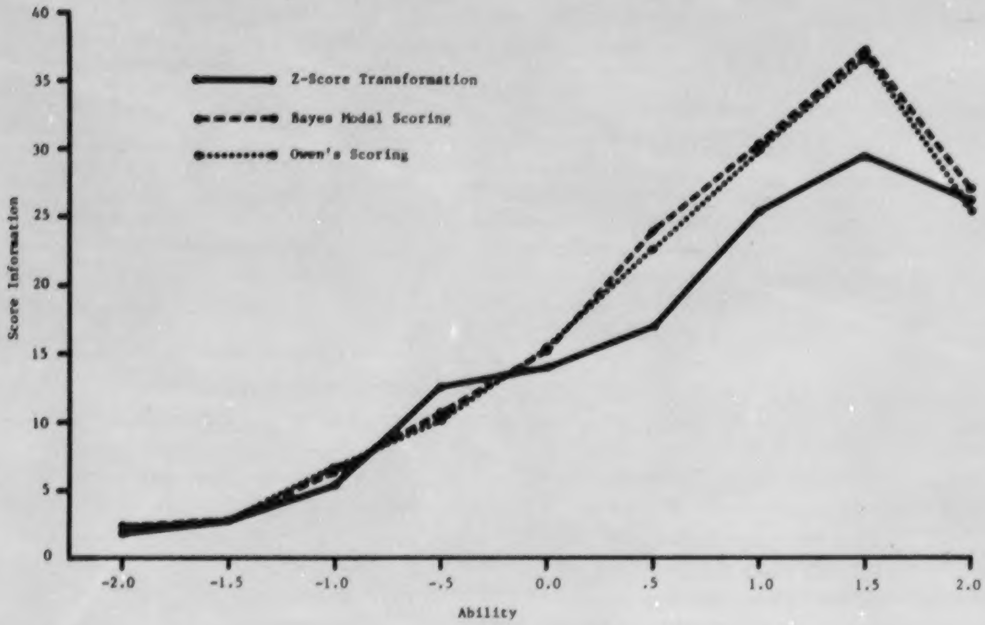


Table 3
Test Score Information of Conventional Z-Score Method and Two Bayesian Scoring Methods Using Estimated and Known Parameters in a Static Test

		Parameters			
		Estimated		Known	
Ability Level	Z-Score	Bayes Modal	Owen´s	Bayes Modal	Owen´s
-2.0	1.910	2.285	2.215	2.185	1.876
-1.5	2.775	2.713	2.817	2.796	3.177
-1.0	5.316	6.343	6.640	6.930	6.963
-0.5	12.545	10.645	10.189	9.986	9.407
0.0	13.963	15.200	15.484	14.807	14.725
0.5	16.876	23.884	22.631	26.225	24.184
1.0	25.307	30.190	29.836	29.441	28.582
1.5	29.386	37.179	36.691	38.875	38.288
2.0	26.066	26.880	25.294	28.293	26.353

the conventional method, at this high level (.94 to .96) the result is highly significant ($p \leq .0001$). The fidelity coefficient of the 80-item VSAT test scored with either Bayesian method is comparable (via the Spearman-Brown proph-

Table 4
Correlation of Known Ability with Ability Estimates for a Conventional Z-Score Scoring Method and Two Bayesian Scoring Methods Using Known and Estimated Parameters in a Static Test

Scoring Method	_r_
Conventional Z-Score Transformation	.941
Bayes Modal	
Estimated Parameters	.959*
Known Parameters	.960*
Owen's Bayesian	
Estimated Parameters	.958*
Known Parameters	.958*

*Values significantly different from conventional Z-score transformation _r_ at $p \leq .0001$.

ecy formula) to a fidelity coefficient of a 120-item test scored conventionally. Also of interest is the fact that the fidelity coefficients computed using either Bayesian procedure with known item parameters were not significantly different from the fidelity coefficients computed from Bayesian scoring with estimated item parameters. This attests to the robustness of the Bayesian scoring procedure to errors in item parameter estimation.

Conclusions. It is apparent that improvements in the measurement of examinees on conventional tests can be realized by the use of mathematical scoring procedures that are based upon latent trait theory. Bias seems to be diminished, and test score accuracy and precision are improved with these two Bayesian scoring procedures, compared to the conventional scoring method.

Study 2:
An Analysis of the Effect of the Correction for Regression and Parameter Estimation Errors Upon Two Bayesian Adaptive Testing Procedures

Purpose

The present study simulated an adaptive test using both Owen's Bayesian procedure and the Bayes modal procedure. The research attempted to determine the effect of item parameter estimation errors upon the test characteristics as a function of ability level. In addition, this study investigated the effect of a correction for regression applied to the ability estimates obtained using the Owen algorithm. The Bayes modal procedure already incorporates this regression correction.

Owen's Bayesian Procedure and the Correction for Regression

The Bayesian adaptive ability estimation procedure has been well documented elsewhere (McBride & Weiss, 1976; Owen, 1975) and will not be reported here. However, to understand the correction, a brief conceptual description is in order. The procedure assumes a normal distribution of the ability estimates with mean 0 and variance 1. The item bank is then scanned to identify the item that will minimize the expectation of the posterior variance of the distribution if administered. That item is then administered, and a new ability estimate (mean of posterior distribution) and variance about that estimate are computed. The ability estimate is then used as the prior mean, and an item is again selected to minimize the expected value of the variance of the posterior distribution. This procedure is repeated iteratively.

A correction for regression is applied to the final ability estimate. The correction consists of dividing the final ability estimate by what Urry (1977) refers to as the test reliability. This reliability is 1.0 minus the Bayesian posterior variance, and this value obviously will differ for each individualized test. Urry believes that more accurate measurement is attained by terminating adaptive tests based on a fixed posterior variance, rather than a fixed number of items. However, Urry (1977) concedes that this correction should be effective for both fixed and variable-length tests. This study investigates the fixed-length test only.

Bayes Modal Adaptive Procedure

The Bayes modal adaptive ability estimation procedure developed for this study consisted of two algorithms--one to estimate ability and one to select appropriate items to be administered. The ability estimation algorithm was based on the Bayesian scoring procedure developed by Samejima, using the item response function and an assumption of a normal distribution of ability. Urry (1976) uses this procedure in the second iterative stage of his item parameter estimation procedure. The item selection procedure chooses that item which provides the highest level of item information for the current ability estimate. The item response function for all administered items is computed. The product of all item response functions and the assumed normal density function is the posterior distribution; the mode of this distribution is the ability level estimate. This value is then unregressed using the same correction as stated earlier. However, unlike the Owen procedure, the corrected estimate is then used as the starting point for the next iteration of item selection.

Design of the Study

Two "ideal" banks were generated, each consisting of 101 items at equal increments of $\underline{b} = .05$ over the range $-2.5 \leq \underline{b} \leq +2.5$. One bank used items whose item discriminations were set at $\underline{a} = 1.6$; the other, at $\underline{a} = .8$. The item parameters were estimated by the ANCILLES program on a group of 50 items based on the responses of 2,000 sims. The procedures differed from Study 1 in that the items were scrambled with the parameters from item banks of another study (Gorman, in prep.). The analysis was based upon three test characteristics as a function of ability level--bias, accuracy, and precison--as documented in Study 1.

Results

Conditional bias. Figure 4 displays the score bias from the 25-item adaptive test employing the Owen algorithm, with and without the correction for regression, and the Bayes modal procedure. The three lines represent the bias in the adaptive procedures using the item bank with item discriminations of a = 1.6, based on estimated parameters. The Owen procedure with the correction provided the least bias.

Figure 4
Effect of Regression Correction Upon Bias of
25-Item Bayes Modal and Owen Adaptive Tests
(a = 1.6) with Estimated Parameters

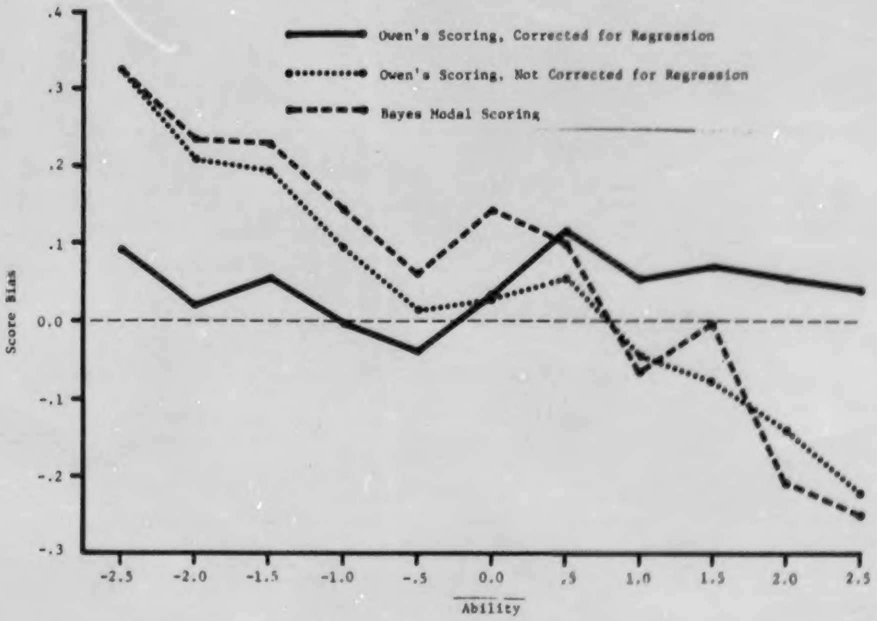


Table 5 shows the effect of regression upon the ability estimates from the Owen procedure using known and estimated parameters. An interesting result is that the regression phenomenon was more prevalent when known parameters were used in the Owen scoring with a correction than with estimated parameters using the same correction. This may be due to sampling errors in parameter estimation working in the preferred direction on this criterion.

Using the less discriminating item bank (a = .8), the regression was more extreme, but the correction using estimated parameters again adequately compensated. The regression correction was less effective when using known parameters.

The Bayes modal adaptive test did not fare as well as the Owen adaptive test. This can be seen in Table 6, which lists the bias for the two item banks under conditions of known and estimated parameters. With known parameters, the bias was tolerable with the better item bank. The bias under the other three conditions was significantly greater.

Conditional accuracy. Table 7 shows the effect of the regression correc-

Table 5
Effect of Regression Correction Upon Bias of the 25-Item Owen's Adaptive Test with Estimated and Known Parameters for Two Item Banks

Item Bank and Ability Level	Parameters			
	Estimated		Known	
	Corrected	Uncorrected	Corrected	Uncorrected
a=.8 Item Bank				
-2.5	.063	.468	-.055	.416
-2.0	.021	.351	-.167	.237
-1.5	-.036	.222	-.127	.179
-1.0	-.029	.146	-.124	.091
-0.5	-.053	.042	-.067	.043
0.0	.008	.007	.008	.006
0.5	.049	-.047	.102	-.019
1.0	.011	-.165	.075	-.143
1.5	-.027	-.283	.154	-.186
2.0	.046	-.306	.268	-.205
2.5	.036	-.403	.275	-.314
a=1.6 Item Bank				
-2.5	.091	.326	-.008	.242
-2.0	.019	.207	-.077	.125
-1.5	.055	.193	.001	.151
-1.0	-.004	.093	-.046	.061
-0.5	-.040	.013	-.068	-.009
0.0	.031	.028	.007	.006
0.5	.115	.055	.071	.010
1.0	.054	-.046	.064	-.050
1.5	.071	-.077	.117	-.059
2.0	.054	-.140	.161	-.077
2.5	.041	-.221	.153	-.150

tion upon the root mean square error (RMSE) of the 25-item Owen adaptive test. The average RMSE value for the Owen ability estimates using known item parameters without the correction was .225; using estimated item parameters with the correction, .233; using known item parameters with the correcton, .241; and using estimated item parameters without the correction, .225.

The a = .8 item bank followed this same trend, only to a greater degree, with the exception that the highest average RMSE value was with the Owen procedure using known item parameters and corrected for regression. This result is counter to the expected result. The reason for this may again be due to errors in item parameter estimation favorable to the Owen procedure. Another trend for both item banks was that the RMSE values were lowest about the mean and increased in magnitude as a function of distance from the mean.

Table 8 lists the RMSE for the Bayes modal adaptive test. On the item bank with a = .8 using estimated parameters, the conditional accuracy was poorer than

Table 6
Bias of the 25-Item Bayes Modal Adaptive Test Using Estimated and Known Parameters, with Two Item Banks

	Item Bank			
	a = .8		a = 1.6	
Ability Level	Estimated Parameters	Known Parameters	Estimated Parameters	Known Parameters
-2.5	.575	.213	.324	.115
-2.0	.382	.133	.232	.085
-1.5	.262	.101	.228	.156
-1.0	.094	.027	.143	.097
-0.5	.049	.035	.058	.047
0.0	.066	.043	.143	.059
0.5	.087	.084	.099	.066
1.0	-.049	-.013	-.067	-.002
1.5	-.067	-.101	-.002	-.038
2.0	-.195	-.092	-.208	-.058
2.5	-.343	-.080	-.251	-.038

Table 7
Effect of Regression Correction Upon Root Mean Square Error of the 25-Item Owen Adaptive Test with Estimated and Known Parameters for Two Item Banks

Item Bank and Ability Level	Parameters			
	Estimated		Known	
	Corrected	Uncorrected	Corrected	Uncorrected
a=.8 Item Bank				
-2.5	.370	.556	.413	.532
-2.0	.423	.497	.454	.417
-1.5	.404	-403	.384	.347
-1.0	.432	.388	.433	.349
-0.5	.368	.305	.387	.310
0.0	.352	.291	.390	.313
0.5	.447	.370	.420	.325
1.0	.401	.372	.445	.376
1.5	.351	.404	.416	.357
2.0	.339	.416	.531	.411
2.5	.395	.512	.541	.475
a=1.6 Item Bank				
-2.5	.2907	.4054	.2440	.3181
-2.0	.2411	.2973	.2481	.2434
-1.5	.2629	.3020	.2202	.2496
-1.0	.1864	.1926	.2088	.1927
-0.5	.2223	.1980	.2353	.2025
0.0	.2297	.2072	.2133	.1906
0.5	.2351	.1945	.2386	.2034
1.0	.2142	.1953	.2399	.2118
1.5	.2069	.1923	.2512	.2060
2.0	.2333	.2452	.2670	.2033
2.5	.2487	.2988	.2861	.2558

Table 8
Root Mean Square Error of the 25-Item Bayes Modal Adaptive Test Using Estimated and Known Parameters, with Two Item Banks

	Item Bank			
	a = .8		*a* = 1.6	
Ability Level	Estimated Parameters	Known Parameters	Estimated Parameters	Known Parameters
-2.5	.783	.445	.589	.393
-2.0	.593	.431	.393	.327
-1.5	.428	.373	.388	.279
-1.0	.366	.386	.290	.245
-0.5	.411	.384	.261	.264
0.0	.412	.351	.339	.290
0.5	.426	.414	.268	.233
1.0	.321	.327	.234	.228
1.5	.328	.395	.172	.215
2.0	.320	.368	.269	.206
2.5	.447	.340	.364	.210

Table 9
Test Score Information of Two 25-Item Bayesian Tests, Using Known and Estimated Parameters, with Two Item Banks

	Item Bank			
	a = .8		*a* = 1.6	
Adaptive Test and Ability Level	Known Parameters	Estimated Parameters	Known Parameters	Estimated Parameters
Owen's Bayesian				
-2.0	2.591	4.738	15.776	17.847
-1.5	5.519	8.359	14.786	22.501
-1.0	5.311	6.475	23.878	21.238
-0.5	7.975	8.726	21.079	21.571
0.0	9.808	9.009	25.752	28.516
0.5	5.181	6.974	26.434	21.809
1.0	5.425	6.021	22.470	21.041
1.5	8.975	9.519	26.369	24.226
2.0	9.863	5.897	18.315	23.473
Bayes Modal				
-2.0	1.325	4.210	5.067	9.898
-1.5	2.850	5.822	5.435	13.496
-1.0	4.476	5.689	8.919	13.241
-0.5	4.859	6.425	11.277	11.670
0.0	6.347	8.906	10.608	12.359
0.5	4.982	5.695	17.014	18.389
1.0	7.565	6.504	24.089	15.992
1.5	6.621	5.594	21.752	19.452
2.0	5.142	7.703	8.609	23.604

with the Owen procedure. On the other hand, with the same item bank using known parameters, accuracy was greater with the Bayes modal procedure. With the better item bank, the Owen procedure was superior to the Bayes modal on this criterion.

Conditional precision. Table 9 lists values of score information for 25-item tests with both Bayesian adaptive methods and two item banks. The item parameter estimation errors rearranged the test score distribution and, hence, its information. The Owen procedure provided more information about the mean and dropped off somewhat at the extremes. The Bayes modal procedure provided considerably less information; hence, the standard error of measurement was larger at all ability levels.

Conclusions

The correction for regression effectively diminished the regression to the mean effect. Fortunately, the errors of parameter estimation provided by ANCILLES worked in favor of less biased measurement. The accuracy of the Owen adaptive fixed-length test with this correction was somewhat poorer with parameters estimated by ANCILLES than with known parameters. This drop in accuracy did not appear to be severe enough to discount the Owen procedure for adaptive testing. The Bayes modal adaptive procedure as implemented in this study needs further work to equal or surpass the Owen algorithm, even with more accurately estimated parameters.

REFERENCES

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1979. (NTIS No. AD A067752).

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A047495).

Bejar, I. I., Weiss, D. J., & Kingsbury, G. G. Calibration of an item pool for the adaptive measurement of achievement (Research Report 77-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1977. (NTIS No. AD A044828)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Croll, P., & Urry, V. W. ANCILLES: A program for estimation of the item parameters of normal ogive and logistic mental test models, in preparation.

Gorman, S. *A comparative evaluation of two Bayesian adaptive ability estimation procedures with a conventional test strategy*, in preparation.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary estimation procedure. In C.L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (PS-75-6). U.S. Civil Service Commission, Personnel Research and Development Center, Washington, DC: U. S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Kolakowski, D., & Bock, R. D. *A FORTRAN IV program for maximum likelihood item analysis and test scoring: Normal ogive model* (Research Memo No. 12). Chicago: University of Chicago, Department of Education, Statistics Laboratory, 1970.

Kolakowski, D., & Bock, R. D. *LOGOG: Maximum likelihood item analysis and test scoring: Logistic model for multiple responses* (Research Memo No. 13). Chicago: University of Chicago, Department of Education, Statistics Laboratory, 1972.

Learmonth, G. E., & Lewis, P. A. W. *Naval Postgraduate School Random Generator Package: LLRANDOM* (Research Report NPS55LW73061A) Monterey, CA: Naval Postgraduate School, 1973.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

McBride, J. R., & Weiss, D. J. *Some properties of a Bayesian adaptive ability testing strategy* (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 1975, *70*, 351-356.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 1969, *34* (4, Pt. 2, No. 17)

Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1977, *1*, 233-247.

Urry, V. W. A monte carlo investigation of logistic model test models (Doctoral dissertation, Purdue University, 1970). *Dissertation Abstracts International*, 1971, *31*, 6319B. (University Microfilms No. 71-9475)

Urry, V. W. Ancillary estimators for the item parameters of mental test models. In W. A. Gorham (Chair), *Computers and testing: Steps toward the inevitable conquest* (PS-76-1). Symposium presented at the 83rd annual convention of the American Psychological Association, Chicago, August 1975. Washington, DC: U.S. Civil Service Commission, Personnel research and Development Center, September 1976. (NTIS No. PB 261 694)

Urry, V. W. Tailored testing: A spectacular success for latent trait theory. Springfield, VA: National Technical Information Service, 1977.

Wood, R. L., Wingersky, M. S., & Lord, F. M. Logist: A computer program for estimating examinee ablity and item characteristic curve parameters (Research Memorandum RM 76-6). Princeton, NJ: Educational Testing Service, 1976.

DISCUSSION: SESSION 1

BRIAN WATERS
AIR UNIVERSITY

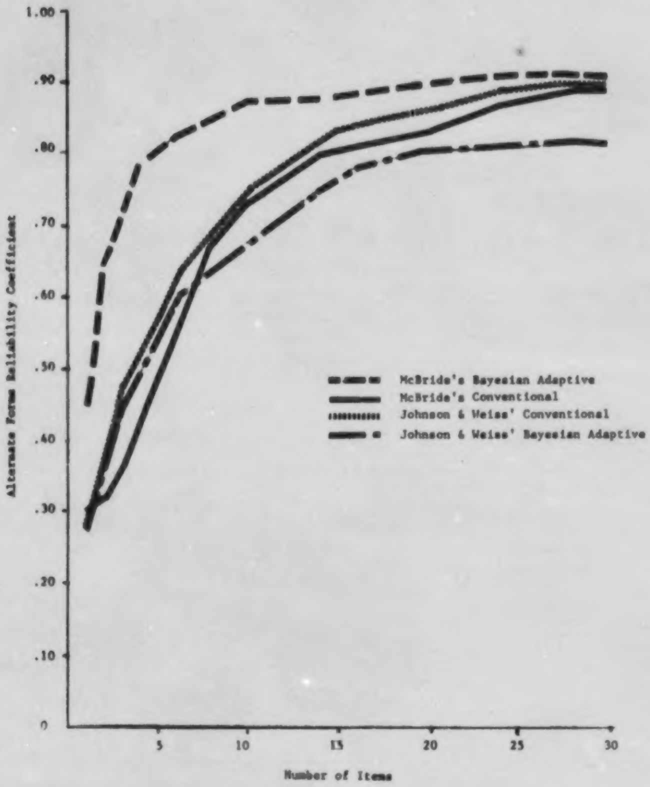
The Department of Defense enlists, classifies, and assigns hundreds of thousands of men and women annually, with test scores a major determinant of these decisions. The testing function must be performed more efficiently, accurately, and equitably; and computerized adaptive testing (CAT) provides the promise of greatly improved large-scale testing efficiencies. Work such as that reported by this session's authors on various adaptive testing strategies is therefore important.

These papers represent two lines of needed research--basic and more applied research on CAT. We still have many theoretical questions best addressed by simulation studies such as Gorman's, as well as myriad practical problems, which are best investigated with live data empirical studies such as McBride's and Johnson and Weiss's. I enjoyed reading each of these papers, particularly the mental exercise of analyzing the contradictory results of the latter two studies.

The primary result from the Johnson and Weiss paper and the McBride paper that caught my attention were the opposite results obtained on McBride's Figure 1 and Johnson and Weiss's Figure 8. These two analyses of Bayesian adaptive testing versus conventional testing both examined parallel forms reliability as a function of test length. McBride's results were consistent with the bulk of similar work done in the past, but the Johnson and Weiss results were startlingly different. The latter paper showed the conventional test yielding consistently higher reliabilities after about 10 items. In an effort to explain this difference in two similarly designed studies, Tom Warm of the Coast Guard Institute, Jim McBride, Marilyn Johnson, Brad Sympson, and I tried to determine what could have led to the conflicting results. Figure 1 shows a plot of the results from the two studies on comparable data. My tentative conclusions attribute the differences to either the parameterization process, the test difficulty, or the examinee characteristics differences. My best "guess" is that the former is the major cause of the contradictory results.

McBride designed an item pool that was extraordinary by any standards. In effect, he followed Urry's guidelines for selection of item characteristics for an adaptive test. All $\underline{a}$ parameters were more than .80 and all $\underline{c}$ parameters were less than .30. His average $\underline{a}$ values for the conventional and adaptive tests were 1.40 and 1.20, respectively. In addition, McBride's items were parameterized on a group of 4,000 examinees from a directly comparable population and produced a nearly rectangular distribution of information.

Figure 1
Alternate Forms Reliability Coefficients
from the McBride and Johnson and Weiss Studies



Johnson and Weiss had test item a parameters as low as .65, with a mean of 1.05 and a range of .65 to 2.25 on the conventional test. The adaptive test a parameter range, however, was .04 to 3.00, with a mean of .76. Particularly in the extreme ranges of ability, some of the items were adding practically no information to the adaptive test. The items were parameterized on far fewer examinees (82 to 1,861 with a median of about 300), and the item distribution was much more peaked. As McBride (paraphrasing Urry, 1970) stated in his paper, "a good tailored test design is superior [to conventional testing], provided that highly discriminatory test items are available." From a purely psychometric viewpoint, I would expect McBride's items to be more effective in an adaptive test as compared to a conventional test and to have more stable item parameter estimates than Johnson and Weiss's.

These contradictory results concern me in another way. Johnson and Weiss's data come from a much more "real world" situation. McBride's careful item selection, parameterization, and design are to be highly commended; however, in many applications, the "ideal" item pool he used is simply just not obtainable. Unfortunately, most of us will be faced with a pool more like that of Johnson and Weiss. If, in fact, their results become typical, the practical application of adaptive testing is threatened. The Johnson and Weiss study thus needs replication.

McBride's study was exceptionally well done. It is nice to see data from the real world rather than from just "Psychology 101" students. I would have liked to have seen test statistics, including reliability, reported for the 50-item criterion test used in the validity analyses. McBride's results of a large increase in reliability with no significant change in validity is not atypical. More information on the criterion measure would have helped the reader conjecture why the validities did not increase with the reliability coefficients. My feeling is that it is related to the fact that the correlation coefficient only uses mean values and that the criterion measure was a conventional test score. If, as the errors of measurement suggest, the adaptive scores had less error variance and more true variance in them, then I would expect less correlation between adaptive and conventional scores than between two conventional scores. The additional true variance would be unique to the adaptive scores, whereas some of the error variance would be common, by chance, to the conventional test scores.

In a recent conversation with McBride, I discovered that since the conference he has acquired another criterion score on the examinees from this study. He reports that the validity coefficients on the adaptive tests were consistently higher (up to .19) than the conventional test validities, with the largest gain at shorter test lengths.

Before leaving these two papers, I would like to comment briefly on McBride's conclusion that fixed test length was as reliable as variable test length. I have a difficult time conceptually accepting this result, if for no other reason than that I believe that individual differences must make a difference. Practically, fixed length is certainly logistically and legally more realistic, which are perfectly valid reasons for using this testing strategy. Theoretically, however, I feel that potential efficiencies must exist with variable length. As Richard Anderson of the University of Illiniois has said, "You can't let bad data ruin a good theory."

Gorman's paper really consisted of two independent monte carlo simulation studies that followed up work suggested by McBride and Weiss (1976) and Urry (1977). It focused on the relative merits of two Bayesian models--the Owen algorithm and Samejima's Bayes modal procedure--and conventional rights-only scoring. Gorman's first study evaluated the efficiency of the two Bayesian models and conventional scoring on static (i.e., nonadaptive, or conventional) tests using three measures of efficiency: (1) average bias, (2) average accuracy, and (3) test score precision. He generated 2,000 simulated examinees (sims) from a normal distribution (mean 0, variance 1) and 80 item scores for each sim for both Bayesian and conventional sim group members. He then used ANCILLES to analyze the data.

Gorman's first study results showed considerably less bias of estimation for the two Bayesian procedures than for the conventional scoring at all points on θ except at θ = -.5 to +.5, with the Owen scoring generally better than the Bayes modal scoring at the lower θ levels and vice versa at the higher θ levels.

On his second measure of efficiency, conditional accuracy, again the conventional scoring yielded less accurate parameter estimation than the two Bayesian methods. Little accuracy differences between the latter two methods evolved, although the Owen procedure did show slightly more error than the Bayes modal model for most of the ability continuum.

Gorman's conditional test score precision measure showed substantial gains for both latent trait scoring models over conventional scoring, with nearly identical results between the two mathematical models. He also found statistically significant, though relatively small (.02), gains in fidelity coefficients in favor of the latent trait models. He concluded from his first study that measurement improvements can be realized through the use of the latent-trait-theory-based models to score static tests.

Gorman's second study was a follow-up of Urry's (1977) suggestion on McBride and Weiss's (1976) study results, which documented the regression to the mean effect using Owen's procedure. Urry suggested dividing the Bayesian regressed ability estimate by the test reliability (the Bayesian posterior variance squared). Gorman followed this procedure in a monte carlo simulation using ANCILLES, the revision of OGIVIA3, for evaluating the efficiency of the Bayes modal and the Owen models with the correction for regression applied.

Gorman's study results showed the Owen procedure to be generally preferable to the Bayes modal procedure in terms of conditional bias, conditional accuracy, and conditional precision when the correction for regression was used.

Considering the work performed on differences between the various computer program ability estimates, such as Bejar and Weiss (1979) showed for different maximum likelihood and Bayesian procedures, I am glad to see studies such as Gorman's being done. Somehow, we need to settle the arguments of the advantages and disadvantages of the various models whereby the results of each study are questioned by the proponents of other models. Algorithm comparisons with known parameters are an effective way to address this research question.

As a final observation on the subject of this session, I was very pleased to see two empirical live-data studies done. Although basic research is important, many of our funding agencies respond more to data from real people as opposed to simulees. I would suggest that future empirical studies include cost data in their battery of dependent variables. There has been a dearth of these data, and they have substantial impact on a funding agency's decisions. I recommend that proposals for future empirical adaptive testing studies should all include cost variables. In the competition for limited research dollars, this information could well be the difference between obtaining funding and not; but more importantly, the information is important for us as adaptive testing researchers.

REFERENCES

Bejar, I. I., & Weiss, D. J. Computer programs for scoring test data with item characteristic curve models (Research Report 79-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1979. (NTIS No. AD A067752)

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

Urry, V. W. A monte carlo investigation of logistic test models. Unpublished doctoral dissertation, Purdue University, 1970.

Urry, V. W. Tailored Testing: A spectacular success for latent trait theory. Springfield, VA: National Technical Information Service, 1979.

SESSION 2*:
ADAPTIVE TESTING IN GERMANY

A VALIDITY STUDY OF AN ADAPTIVE TEST OF READING COMPREHENSION

LUTZ F. HORNKE AND
MICHAEL P. SAUTER
UNIVERSITY OF DUSSELDORF

COMPUTERIZED TESTING IN THE FEDERAL ARMED FORCES

WOLFGANG WILDGRUBE
GERMAN MINISTRY OF DEFENSE

---

*A paper entitled "Test Construction Using Adaptive Administration Techniques," by Wolfgang Buchtala, was also presented in Session 2, but was not available for inclusion in these Proceedings.

A VALIDITY STUDY OF AN ADAPTIVE TEST OF READING COMPREHENSION

LUTZ F. HORNKE AND MICHAEL P. SAUTER
UNIVERSITY OF DÜSSELDORF

Adaptive means that a test adapts to the testee's proficiency level in the proper "can-do" sense. A fair number of items are placed at the testee's disposal; and solely by means of tactical rules, the testees self-select their own individual subset of items. To achieve this, their previous responses are used to help in making item-to-item decisions. In addition, restrictions on test time are imposed to insure unidimensional interpretations.

In the literature many variants of adaptive schemes are described and discussed (see Hornke, 1976, 1977, 1979a, 1979b, 1979c; Hornke, Sauter, Suessmilch, & Burghoff, 1979; Lord, 1971; Weiss, 1974; 1975; Weiss & Betz, 1973; Wood, 1973). Generally speaking, the idea utilized is that of branching from item to item or between groups of items utilized: The item someone is branched to is made contingent on his/her response(s) to earlier item(s). Thus, whenever a testee answers an item correctly, he/she is presented with more a difficult one on the assumption that his/her proficiency level at this intermediate stage is somewhat higher than that displayed in the item just mastered. The contrary holds for incorrect responses. The complexity and variety of branching rules is not limited (see Hornke, 1976; Weiss, 1978). The more flexible the branching technology, the more adaptive the decision process will be, and this yields very reliable information about a testee's proficiency and his/her can-do potential.

The term branching technology is used here intentionally because many adaptive testing projects already use computers. According to highly sophisticated estimation procedures based on probabilistic mathematical response models (see Fischer, 1978; Lord & Novick, 1969), items are deliberately retrieved from a larger pool. These approaches use item parameters to estimate a person's probable standing. After several cycles of item administration and parameter estimation, a person parameter emerges that confidently reflects an individual's proficiency level. Since items and persons are calibrated on the same scale, by looking at those items (i.e., behaviors), the parameters of which lie in the vicinity of the person parameter, interpretations are readily available.

Computer terminals and micro-computers are quite costly, however, so that paper-and-pencil versions deserve some attention. The basis of their measurement is somewhat less stringent compared with flexible computer-assisted tests; but when properly designed, they should allow equivalent or even better measurement precision than conventional tests (see Hornke, 1979b, Hornke et al., 1979).

The test booklet may look the same as that for conventional tests; the difference is that the testee is asked to use a special pen for marking his/her answer. He/she has to pass this lightly over a bracketed field next to the chosen answer. Chemicals then react and render visible the number of an item to be attempted next. By following these numbers as they appear, a testee is branched through the item set (see Hornke, 1979a; Sauter, 1978, 1979; Sauter & Hornke, 1979). The testee is intended to be guided to just that subset of items that tells something about his/her can-do level, while leaving out all the other boring or otherwise frustrating items. Since a testee zig-zags through a pyramidal item arrangement, he/she will finally end in a score category, a self-evaluating feature of this tactical test design.

Thus, with branching tactics, flexible, fair, self-scoring, and interpretable tests are at hand. Since any mathematical response model or pyramidal pencil-and-paper test rests, respectively, on the quality of the items and the model or arrangement more successful assessment is guaranteed as long as quality levels are maintained. Even conventional tests, however, require some degree of item validity and reliability, unless any interpretation is better than random guesswork. Whether and how adaptive tests will and should be used is still an open research question.

Adaptive Test Designs

Individualization is a concept that meets approval on many different sides. To some extent, assessing an individual in his/her own right solely by what he/she is doing seems fair. Saving time by asking nonsuperfluous questions capitalizes more on the economy and less on the psychology of testing, though in that area, too, something might be gained. Reduction of the stress induced by testing, maintenance of motivation, and lack of boredom are but a few psychological effects. So far very little is known about these side effects and the benefits of individualized testing; these seem to be areas of potential that await further evaluative research.

At present, individualized testing is thought to have positive or at least non-negative effects on testees. To understand the entire range of adaptive programs better, three possible adaptive designs are considered below.

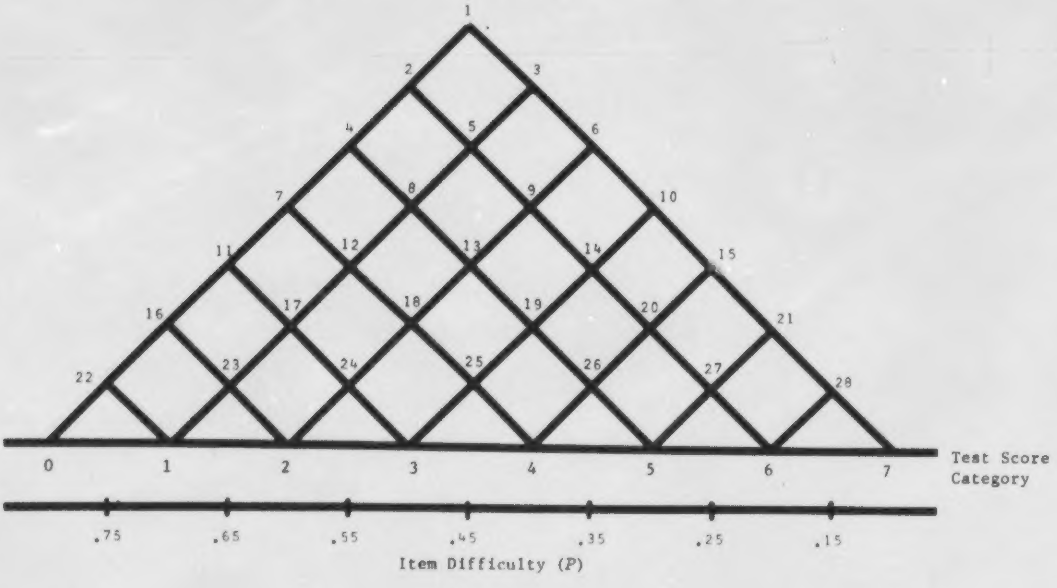
Curtailed item sampling. This approach, a naive type which has some intuitive appeal, resembles the examination models used in classrooms. A teacher asks a student several questions, with content and complexity varying according to the answers given. After a specified period of time the teacher stops and evaluates the student. In comparing several oral examinations, considerable variation would easily be found in the number as well as in the difficulty of questions: This is a genuinely adaptive approach. Thus, two students may earn the same grade but may have been asked different questions as far as number and/or complexity was concerned. Variation in the number seems fair because students who are asked more have a chance to demonstrate their true behavior level; whereas with others, final evaluations are quite obvious after only a few questions.

Computer-assisted testing. Curtailing the numbers of questions, i.e., restricting the sampling of items from a behavioral domain, is a reasonable decision. For adaptive tests this would mean evaluating a testee's distance from a set of criterion levels. Testing is stopped when, for a fixed number of items, a testee is irrevocably located on either side of the decision point. This may be achieved fairly soon. When there are 16 items and the criterion is set at 50%, testing should stop after 8 correct responses. However, this could occur when all the first 8 responses are correct. A testee who made an error on the first 2 items has to be tested for at least 8 more items, yielding a total of 10. Varying numbers of items will occur when several students are tested, one typical aspect of flexible adaptive tests. The example of an oral examination given above dealt with two possible adaptation criteria: (1) the number of questions before a terminal decision can be made and (2) the quality of questions needed to make a procedural decision. A very flexible adaptive testing program will have to consider both criteria; this may be possible with computer-assisted testing (see Weiss, 1975, 1978).

Paper-and-pencil pyramidal tests. Since large-scale adaptive testing by means of computers is hampered by costs, other means have been invented and used to achieve a branching test system, even with group testing; and a pyramidal test design for use with paper-and-pencil devices has emerged. According to its feasibility and overall value, it lies somewhere between curtailed sampling and computer-assisted testing. By *pyramidal* is meant an item arrangement that is structured like a network. For a certain population the item locations on some dimensions are known.

In order to design such a test, items are deliberately selected to form a desired hierarchical item order (see Figure 1). At the top the testee gets the starting item (Item 1), which has to be answered by each candidate. When a cor-

Figure 1
Model of a Pyramidal Item Order



rect answer is given, testees are branched to the right. Consequently, a more difficult item has to be attempted. The contrary holds for an incorrect response. Thus, contingent on their responses, testees are individually branched through the item arrangement and will finally end in a test score category that tells something about the behavioral level attained.

To contrast this approach with curtailed and computer-assisted testing, it becomes quite obvious (1) that there are available far more items than a given testee has to attempt, (2) that testees find their individual paths through the item network and come (or ought to come) close to the upper bounds of their proficiency level, (3) that testing ends after a preset number of items has been attempted, and (4) that the final item leads directly to a test score category, i.e., no further scoring is necessary because the test is essentially self-scoring. The dominant design feature is to adapt the quality of the items, and not their number, to any testee.

The pyramidal test is a fixed strategy as far as item number and arrangement are concerned, but a testee works more or less flexibly on items that are assumed to suit his/her proficiency level more and more. The technical problems with the pencil-and-paper format and group testing were undertaken by means of chemicals. The list of adaptive test designs here is far from complete; many other versions have been described (e.g., see Hornke, 1976; Weiss, 1976, 1978). The report above was meant to examine closely various construction characteristics, i.e., flexibility in item number, item difficulty, or both.

Construction of an Adaptive Pyramidal Test

The studies of Sauter (1978) and of Hornke et al. (1979), looked closely at the adaptive test format and especially at the pyramidal item order in use. It was the aim of both Sauter (1978) and Hornke et al. (1979) to construct and to evaluate such a test design; nevertheless, the choice of the linguistic item material was not accidental. The pyramidal item order requires question forms that can be evaluated objectively, e.g., multiple-choice items or items with a blank. Moreover, it should be possible to rank these items according to their empirical, as well as according to their content, difficulty, which should reflect a higher level of linguistic competence. In addition, the choice of the item material was influenced by the fact that it was not possible, or necessary for this purpose, to construct and to evaluate new items. It was therefore inevitable to seek proven items in existing tests.

One test that approximately meets the above prerequisites is the Cologne Placement Test (see Bonheim & Kreifelts, 1979), which is a traditional placement test for students at the beginning of their first semester in the course "English as a Foreign Language." It consists of four subtests: Vocabulary, Grammar and Usage, Reading Comprehension, and Style and Verbal Logic. According to the needs of a pyramidal item order, reading comprehension items seemed to fit best.

In fact, however, it is not very easy to show what reading comprehension questions actually do test. Definitions are usually tautological: "Reading comprehension tests the ability to read and to understand a particular language." This definition, however, covers a multitude of aptitudes that have only been described very incompletely up to now. Some language and test experts (see Harris, 1969; Heaton, 1975; Lado, 1967; Pynsent, 1972) have tried to discover a few of the factors involved and to put them into a hierarchical order with regard to their level of difficulty and complexity. Obviously, at a more basic level reading comprehension requires the understanding of the meaning of words or word groups in the context in which they appear as well as the recognition of structural clues and the comprehension of structural patterns. These aspects of language are usually dealt with in tests of vocabulary and grammar--that is, the testee has to show his/her ability to ascertain the verbal meaning of a straightforward sentence or phrase. On an advanced level, reading comprehension involves higher mental abilities, such as how to comprehend paragraphs and to select the main ideas, how to draw conclusions from the text, and how to make inferences and to read between the lines. The level of reading comprehension that is actually tested depends to a certain extent on the item type that is used. For example:

Example 1

He asked me to ...... him two thick slices of beef.
(A) carve (B) slash (C) peel (D) split (E) shave
(Jackson, 1976, p. 171)

It is obvious that this question form does not put too great a demand on the testee's reading comprehension abilities and can rather be looked upon as a vocabulary item. The testee has only to know that "carve" is the appropriate word for meat. He/she can answer this item correctly just on the basis of his/her knowledge of vocabulary. To a limited degree this item type can also test grammatical knowledge by offering choices/words that all seem to fit according to their meaning; but, in fact, only one fits for syntactical reasons. With this item type it is therefore very difficult to say to what extent reading comprehension is involved (cf. Jackson, 1976).

Item types that do not lay too much stress on the knowledge of particular words are more usual, and items consisting of a short reading extract of only a few sentences that ask the testee to interpret it in some way seem more appropriate.

Example 2

Parents can give their children enormous help so long as they don't talk too much, give the game away, or block the children's thought. "Come along, dear, we're going to play with this lovely clay, let's see what we can make with it. I think we can make a lovely elephant, come along, what about the trunk dear..." That poor child will have made a mental note that whatever he takes up as a career it won't be sculpture.

Why is this child called "poor"?

(a) He is not allowed to work out his own ideas.
(b) He will never wish to become a sculptor.
(c) He has begun to dislike playing with clay.

(d) He is being taught skills for which he is too young.
(Sauter & Hornke, 1979, p. 165)

Example 2 shows clearly that it tests not only the testee's knowledge of syntactical structures and vocabulary but primarily his/her ability to interpret the text in some way, for the correct answer is not just a paraphrase of the item stem. This item type seems to be capable of testing what Carroll (1968) calls "complexity of information processing--at what level of complexity can the individual process linguistically-coded information?" (p. 53)

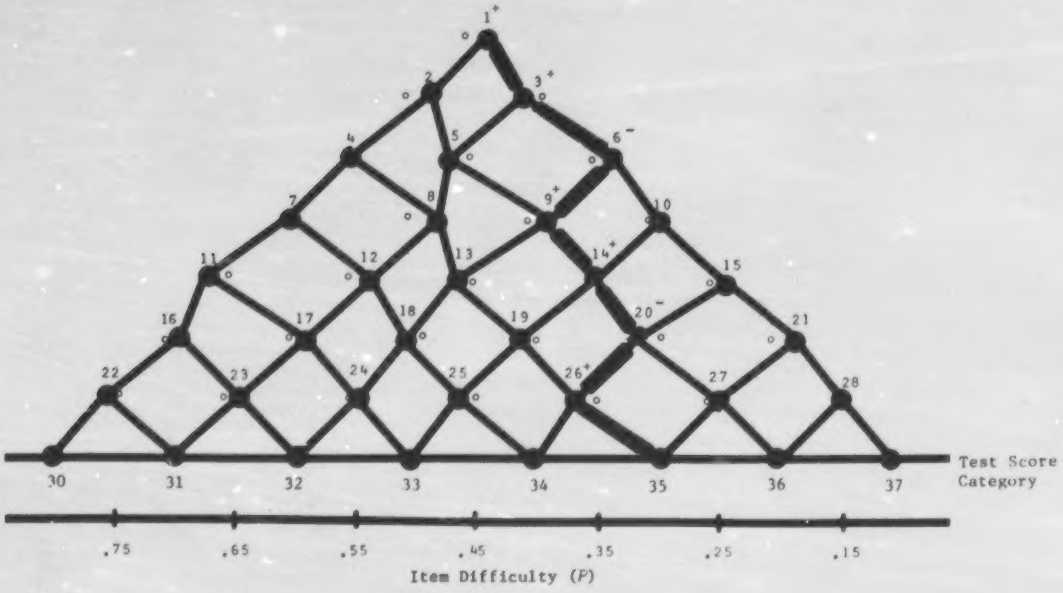
This should be the linguistic dimension that reading comprehension items test, at least in the adaptive test. In practice, however, it is very difficult to find items that represent this dimension even approximately. Even factor-analytic studies can give little help. Thus, it is inevitable that what were regarded as reading comprehension items in the above-mentioned sense do, in fact, correspond rather to a lower level of reading comprehension. The problem with any test construction is that this can cause some confusion, especially in the pyramidal item order by branching testees to incorrect items with regard to their own level of reading comprehension ability.

In this Cologne Placement Test (Bonheim & Kremfelts, 1979), reading comprehension items had been administered to an average of 750 students (up to a maximum of nearly 2,000 students) from 1974 until 1978. Since the placement test had been newly assembled at the beginning of each semester, proven as well as newly constructed items were used, and those items that did not turn out to be satisfactory were left out. The item pool finally contained 88 items from which items were systematically borrowed in order to construct the adaptive test. Each of the 88 items had been carefully analyzed to see whether it could be placed at a certain branching point within the pyramidal item order (see Figure 1). However, with the present state of knowledge, these decisions were not easily made, because there were neither guidelines nor previous experience for item selection that could guarantee a successful branching order. Additional problems that had to be solved were those of time limits and the positional effects of the items in the Cologne Placement Test (for a detailed description, see Hornke et al. 1979; Sauter, 1978; Sauter & Hornke, 1979).

Twenty-eight items were borrowed from the item pool in order to form a pyramidal test, which consisted of seven stages and extended to a difficulty level from P (Probability of a Correct Response) = .75 to P = .15. All items were placed on branching positions according to their empirical difficulty and discrimination. Figure 2 compares the ideal item order with the actual order that is based on the available item data. It shows only relatively small deviations from the positions on the ideal model.

The testee begins with a medium-difficult item ($P_1 = .45$) and is branched to a more difficult ($P_3 = .40$) or an easier item ($P_2 = .50$), depending on whether he/she answered the preceding item correctly or incorrectly (see Figure 2). In this way, he/she is branched through the item order until he/she finally reaches his/her score group. He/she is given only one item at each stage, which eventually means that he/she has to work on only 7 out of 28 items. This seems to be reasonable, assuming that those items that are easier than the items he/she an-

Figure 2
Pyramidal Order of the Adaptive English Test with
the Branching Path of a Hypothetical Testee



swered correctly are probably too easy for him/her. On the other hand, those items that are more difficult are supposedly too difficult for him/her; he/she would most probably answer them incorrectly (see Hornke, 1976, 1977). Thus, only those items are presented to the testee that are most suited for him/her using the pyramidal item order. With the test under consideration, the invisible ink response mode was used in a group setting.

Results of Two Empirical Investigations

Two adaptive reading comprehension tests were investigated--one in a pilot study by Sauter (1978) and the other in a larger validity study by Hornke et al. (1979). Both studies showed that adaptive testing by means of the paper-and-pencil version is quite feasible in group settings. Students had hardly any problems in following branching instructions properly by themselves.

Validity of the Pyramidal Item Order

The design of Sauter's (1978) study asked each student (1) to work through an item set of 28 items in the branching manner and (2) to solve all items left out during the branching in the conventional manner. This yielded two scores per person--an adaptive score and a conventional score, where the first was based on 7 items and the second on 21 residual items. Thus, complete response data were available on all items. This allowed the validity of the pyramidal item order to be investigated in some detail.

The results of an item analysis indicated that all 28 items had become easier than in the original conventional test. However, rank orders between previous and present item difficulties correlated as highly as r = .77, indicating that the order as such had largely survived. Of particular interest was the correlation between scores on the 7 adaptive items and the 21 conventional remaining items, which was r = .47 for 93 testees. Taking the unreliability of the entire set of items into account, however, a stepped-up correlation of r = .64 resulted. Thus, a score based on the 7 optional items had quite a reasonable predictive power to a score based on 21 items.

Validity of the Adaptive Test

The second study (Hornke et al., 1979) had two main purposes, namely, to investigate the validity of an adaptive test and to look at the details of pyramidal item hierarchies. In order to answer the first question, a multitrait approach was used. According to the underlying theory, reading comprehension items ought to call for processes that are different from vocabulary or grammar exercises. Thus, it was expected that there would be a closer relationship between scores for an adaptive and a conventional reading comprehension test than with scores from both grammar and vocabulary tests. The study used a two-method--Adaptive versus Conventional--by three-trait--Reading Comprehension (RC) × Grammer (G) × Vocabulary (V)--design. Due to financial restrictions, however, it was impossible to investigate adaptive and conventional test formats with all three traits. The study thus contrasted adaptive versus conventional reading comprehension only.

It is quite obvious that all three traits should correlate with each other because they are genuine parts of language behavior themselves. However, the results in Table 1 indicate that despite all that they have in common, the three item sets measured quite differentiable aspects that pertain to the hypothesized discriminant relation. This means, too, that the data warrant an interpretation of three different traits, even though intercorrelations were not zero (but they are low enough).

However, reading comprehension scores, assessed either in the adaptive or in the conventional way, did not converge to the extent expected. The resulting correlations were too low for tests designed to measure the same trait. The correlation between RC_1 (adaptive) and RC_1 (conventional remainder) especially contradicted any convergent interpretation, despite the fact that both item sets are virtual subsets of a larger one. Here, a correlation of .6 to .7 would be more suitable to justify any convergence. It still remains an open question whether adaptive branching of items used with reading comprehension tests introduced a source of error or variation that accounted for the low correlations. A comparison of RC_1 (conventional remainder) with the RC_2 (conventional) scores indicates some dissimilarity in the item sets, which appear to be more different than their common label would lead one to expect.

Conclusions

Although adaptive tests are initially intriguing, there are many problems to overcome. The major problem lies in the fact that for foreign language testing, a properly defined construct is necessary. Consequently, all items ought

Table 1
Correlations Between All Tests and Formats Used

	RC_1		Conventional (28 Items)		
Variable	Adaptive (7 Items)	Conventional Remainder (21 Items)	RC_2	G	V
Convergent					
RC_1					
Adaptive	-	.405	.379		
Conventional Remainder	.531	-	.419		
RC_2					
Conventional	.218	(.403)	-		
Discriminant					
G (Conventional)	.295	.511	(.068)	(.419)	
V (Conventional)	.355	.431	(.214)	-	

Note. Correlation coefficients in parentheses are based on group means instead of individual data.

to belong to an appropriately defined behavioral domain. This is not always easy to achieve, and there might often be a lack of expert consensus. Instead, empirical studies are needed to substantiate any item's relation to the construct in question.

A quite substantial problem for adaptive tests may be seen in the necessary heierarchical order for a pyramidal arrangement. Any branching decision here implies strongly that the hierarchy is valid and stable across samples of the population. The two studies cited above indicated, however, that this may not be the case. As far as there are changes in item difficulties from one sample to the other, this might not matter very much as long as all item positions stay within the hierarchical order intended. Whenever there are changes or shifts in positions, the pyramid is invalidated locally, and false branching occurs. To circumvent this problem, rigorous item analysis may help to keep this weakness within limits. It has to be questioned, too, whether difficulty indices (i.e., the proportions of answers correct) are good and reasonable criteria for a hierarchical ordering of items. With narrowly defined populations and applications, this might be practicable. However, better estimates of an item's scale and hierarchical position are available and should be used. With these two studies cited, it was not possible to perform item analyses, since data were not available for this purpose.

Taking these two arguments together, it follows immediately that there will be hardly any chance to take a conventional test, to rearrange its item order, and to get an adaptive version. With any test construction, careful item writing and analysis is necessary. This is true for adaptive as well as conventional tests; ad hoc test construction hardly conforms to the careful scrutiny that is called for. It should not be expected that adaptive or conventional tests from this source have any value in decision making at all. In foreign language testing only after a good deal of research and empirical investigation has been carried out will there be adaptive tests for a variety of purposes; but, in fact, they are essential in a program where students' proficiency is expected to vary considerably and where decisions of some kind are to be made.

REFERENCES

Bonheim, H., & Kreifelts, B. *Ein universitätseingangstest für neuphilologen abschtussbericht der arbeitsgruppe sprachtests (as) an der Universität Köln zur Verlage beim BMBW*. Köln: Universität Köln, 1979.

Carroll, J. B. The psychology of language testing. In A. Davies (Ed.), *Language Testing Symposium*. London: Oxford University Press, 1968.

Fischer, G. H. Probabilistic test models and their applications. *German Journal of Psychology*, 1978, *8*, 298-319.

Harris, D. P. *Testing English as a second language*. New York: McGraw-Hill, 1969.

Heaton, J. B. *Writing English language tests*. London: Longman, 1975.

Hornke, L. F. *Grundlagen und probleme antvortabhängiger testverfahren*. Frankfurt: Haag & Herchen, 1976.

Hornke, L. F. Antwortabhängige testverfahren: Ein neuartiger ansatz psychologischen testens. *Diagnostica*, 1977, *23*, 1-14.

Hornke, L. F. Four realisations of pyramidal adaptive testing. *Programmed Learning and Educational Technology*, 1979, *16*, 164-169. (a)

Hornke, L. F. Konstruktion eines adaptiv-antwortabhängigen fragebogens zur erfassung von preufungsangst. *Diagnostica*, 1979, *25*, 208-218. (b)

Hornke, L. F. *Testdiagnostische untersuchungsstrategien*. In K.-J. Groffman & L. Michel (Eds.), *Handbuch der psychologischen diagnostik* (Vol. 6, 2nd ed.). Göttingen: Hogrefe, 1979. (c)

Hornke, L. F., Sauter, M. F., Süssmilch, B. H., & Burghoff, U. R. *Konvergente und diskriminante validität eines adaptiv-antwortabhängigen Englischtests fur Anglistikstudenten* (DFG HO-758-1). Unpublished research report, Universität Düsseldorf, Erziebungswissenschaftliches Institut, 1979.

Jackson, S. H. Reading comprehension questions in tests of English as a foreign language. In *Kongressberichte der 7. Jahrestagung der Gesellschaft für Angewandte Linguistik*. Tier: GAL e.V., 1976.

Lado, R. *Language testing*. *The construction and use of foreign language tests* (5th ed.). London: Longmans, 1967.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Lord, F. M. The self-scoring flexilevel test. Journal of Educational Measurement, 1971, 8, 147-151.

Michel, L. Allgemeine grundlagen psychometrischer tests. In R. Heiss, K.-J. Groffmann, & L. Michel (Eds.), Psychologische diagnostik. Handbuch der psychologie (Vol. 6). Göttingen: Hogrefe, 1964.

Popham, W. J. The case for criterion-referenced measurement. Educational Researcher, 1978, 7, 6-10.

Pynsent, R. B. The objective reading comprehension test. In R. B. Pynsent (Ed.), Objektive tests in Englishchunterricht der schule und universität. Frankfurt: Athenaeum, 1972.

Sauter, M. P. Entwicklung und erprobung eines antwortanhängigen testverfahrens zur überprüfung des leserverständnisses in Englisch. Unpublished doctoral dissertation, Universität Düsseldorf, Erziehungswissenschaftliches Institut, 1978.

Sauter, M. P. Adaptive tests im Fremdsprachenunterricht. In Kongressbericht der 9. Jahrestagung der Gesellschaft fur Angewandte Linguistik (Vol. 3). Heidelberg: Julius Groos, 1979.

Sauter, M. P., & Hornke, L. F. Adaptives testen im Englischunterricht. Entwicklung eines flexiblen testverfahrens zur messung von leserverständnis. Anglistik & Englischunterricht, 1979, 8, 151-166.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

Weiss, D. J. Computerized adaptive trait measurement: Problems and prospects (Research Report 75-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, November 1975. (NTIS No. AD A018675

Weiss, D. J. Computerized ability testing 1972-1975 (Final Report). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1976. (NTIS No. AD A024516)

Weiss, D. J. Proceedings of the 1977 Computerized Adaptive Test Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? (Research Report 73-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, February 1973. (NTIS No. AD 757788)

COMPUTERIZED TESTING IN THE FEDERAL ARMED FORCES

WOLFGANG WILDGRUBE
GERMAN MINISTRY OF DEFENSE

The Federal Armed Forces (FAF) consists of about 480,000 soldiers (240,000 of these are draftees); the FAF administration comprises 170,000 civilians; and in the FAF Psychological Service there is a civilian staff of 1,300 psychologists. Figure 1 presents an overview of the organization of the FAF Psychological Services. The center of activities is in personnel psychology, with more than 80% of the psychologists in the area of aptitude diagnosis. Figure 2 shows the psychological aptitude testing procedures for selection and classification for both the FAF and the FAF administration. Aptitude diagnoses are carried out for various purposes for large samples, such as for draftees (about 300,000 diagnoses per year); for volunteers (about 30,000 per year); for advancement from sergeant to an officer career; and for selection of pilots, pyrotechnists, civil servants, and personnel for linguistic services. Aptitude and intelligence tests are administered by paper and pencil to groups of about 50 persons. Special apparatus tests or other special procedures and psychological interviews follow as necessary, dependent on the selection process or on the individual result. With these procedures, the Psychological Service thus attempts to make the best possible personnel decision.

Problems

The large number of testing procedures and the wide areas of testing create numerous problems. Mass testing (about 350,000 testees per year) requires a large quantity of material and manpower. The test application, scoring, and decision-making consist of many routine activities that require a great expenditure of personnel.

For each selection procedure all testees of a group process standard testing batteries: All testees undergo the same test battery during a limited period of time. For a certain number of testees the test is too difficult; for others, too easy. Thus, motivation decreases and fatigue increases. Special knowledge, attitudes, or personal spheres of interest or inclinations are not taken into consideration. Moreover, very rarely are special procedures possible, so that in the limited time allotted only some aptitude dimensions are carried out in an undifferentiated manner.

At present the mass data, collected by paper and pencil, do not permit follow-up analyses. Statistical evaluations of the testee data are impossible, and

Figure 1

Organization of the Psychological Services in the Federal Armed Forces of Germany

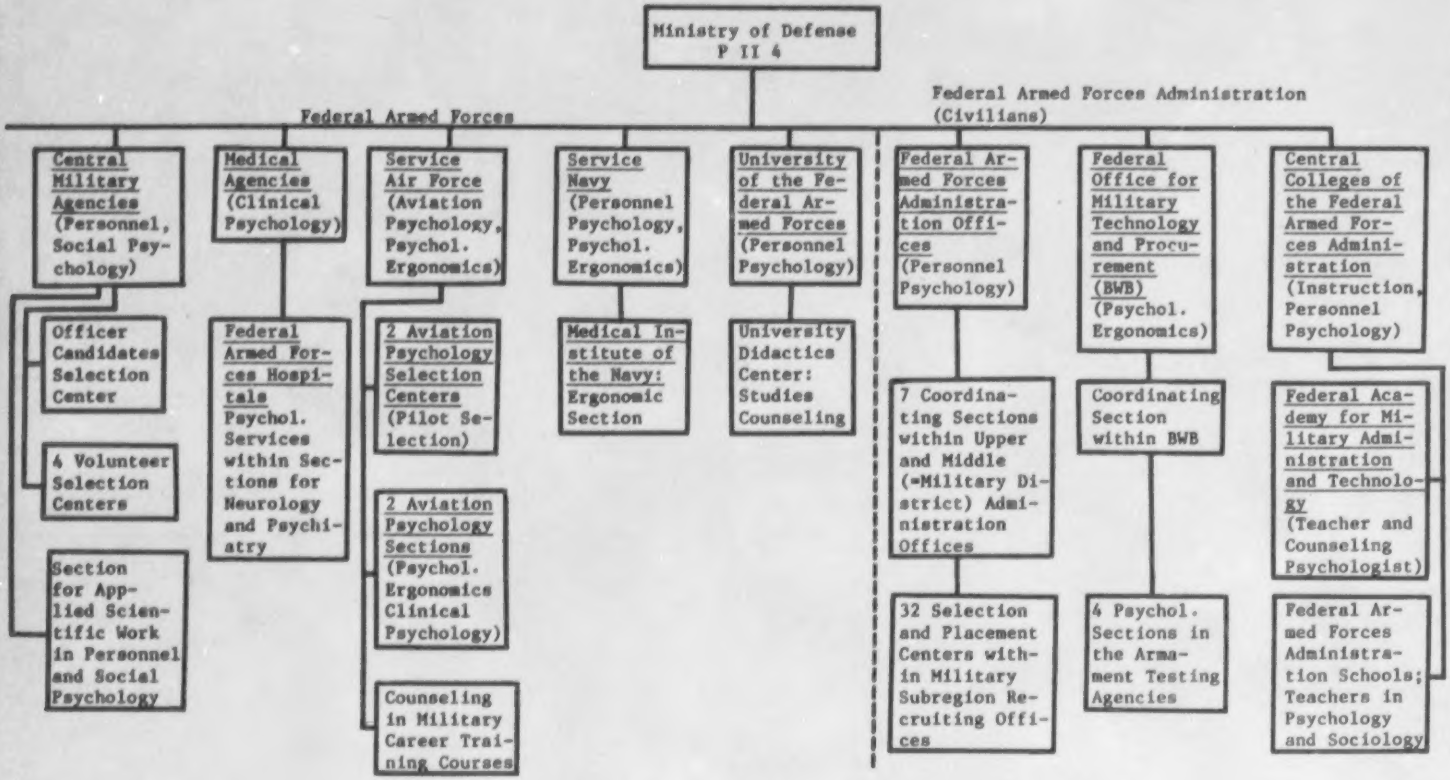
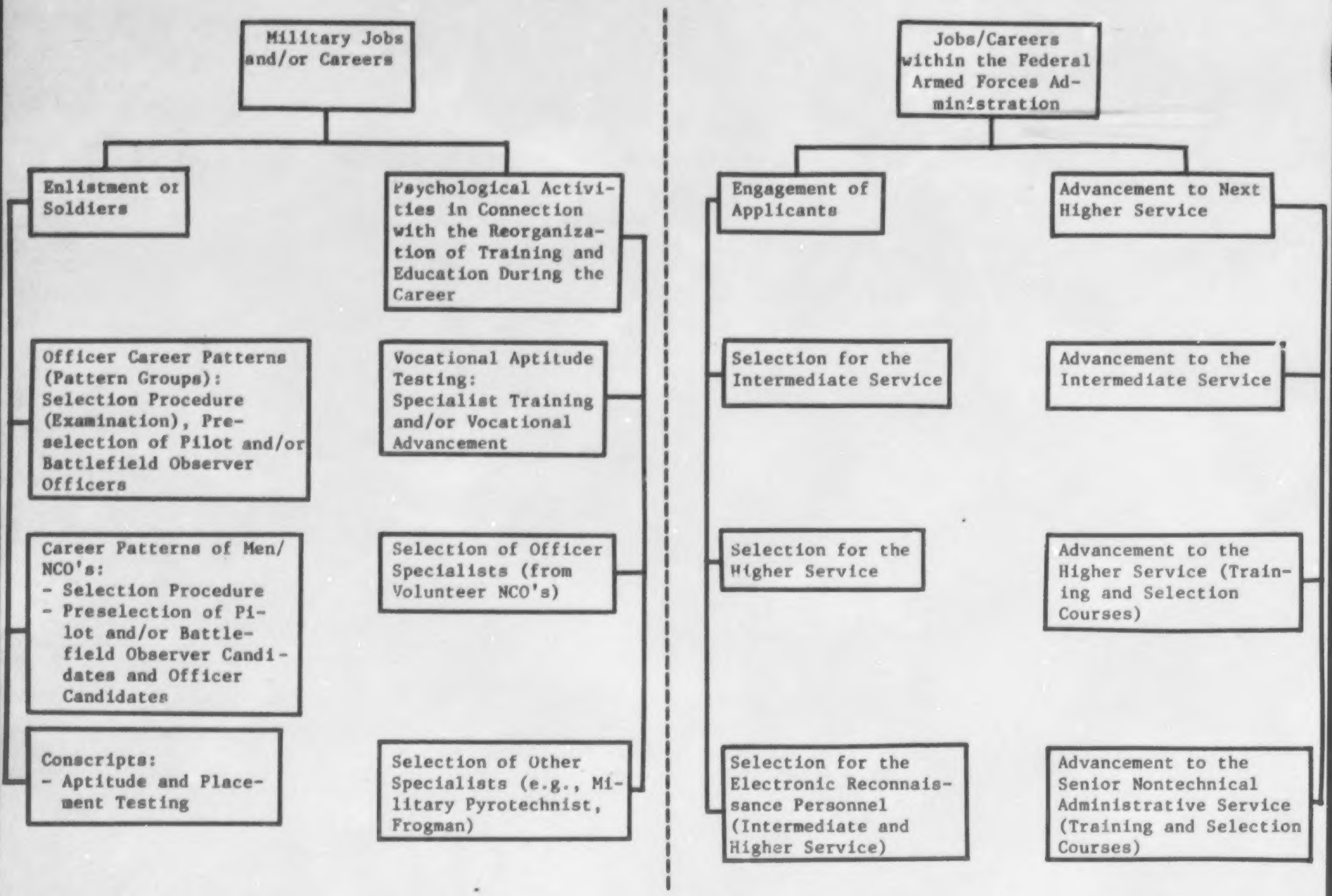

Ministry of Defense P II 4
Federal Armed Forces
Federal Armed Forces Administration (Civilians)
Central Military Agencies (Personnel, Social Psychology)
Officer Candidates Selection Center
4 Volunteer Selection Centers
Section for Applied Scientific Work in Personnel and Social Psychology
Medical Agencies (Clinical Psychology)
Federal Armed Forces Hospitals Psychol. Services within Sections for Neurology and Psychiatry
Service Air Force (Aviation Psychology, Psychol. Ergonomics)
2 Aviation Psychology Selection Centers (Pilot Selection)
2 Aviation Psychology Sections (Psychol. Ergonomics Clinical Psychology)
Counseling in Military Career Training Courses
Service Navy (Personnel Psychology, Psychol. Ergonomics)
Medical Institute of the Navy: Ergonomic Section
University of the Federal Armed Forces (Personnel Psychology)
University Didactics Center: Studies Counseling
Federal Armed Forces Administration Offices (Personnel Psychology)
7 Coordinating Sections within Upper and Middle (=Military District) Administration Offices
32 Selection and Placement Centers within Military Subregion Recruiting Offices
Federal Office for Military Technology and Procurement (BWB) (Psychol. Ergonomics)
Coordinating Section within BWB
4 Psychol. Sections in the Armament Testing Agencies
Central Colleges of the Federal Armed Forces Administration (Instruction, Personnel Psychology)
Federal Academy for Military Administration and Technology (Teacher and Counseling Psychologist)
Federal Armed Forces Administration Schools; Teachers in Psychology and Sociology

Figure 2
Psychological Aptitude Testing in the Federal Armed Forces of Germany


Military Jobs and/or Careers
Enlistment of Soldiers
Officer Career Patterns (Pattern Groups): Selection Procedure (Examination), Preselection of Pilot and/or Battlefield Observer Officers
Career Patterns of Men/NCO's: - Selection Procedure - Preselection of Pilot and/or Battlefield Observer Candidates and Officer Candidates
Conscripts: - Aptitude and Placement Testing
Psychological Activities in Connection with the Reorganization of Training and Education During the Career
Vocational Aptitude Testing: Specialist Training and/or Vocational Advancement
Selection of Officer Specialists (from Volunteer NCO's)
Selection of Other Specialists (e.g., Military Pyrotechnist, Frogman)
Jobs/Careers within the Federal Armed Forces Administration
Engagement of Applicants
Selection for the Intermediate Service
Selection for the Higher Service
Selection for the Electronic Reconnaissance Personnel (Intermediate and Higher Service)
Advancement to Next Higher Service
Advancement to the Intermediate Service
Advancement to the Higher Service (Training and Selection Courses)
Advancement to the Senior Nontechnical Administrative Service (Training and Selection Courses)

changes within the tests--for singular items or for the whole norm values--are not analyzed. Technical, organizational, and legal problems (as for instance, the security of tests) are connected with the mass testing and the different areas of aptitude diagnoses. It is necessary that the tests and the selection procedures be modified shortly. Above all, not only do the tests--that is, the psychological selection procedures within the scope of decisions or careers--become obsolete very soon but the patterns for their solutions (the items and the corresponding correct answers) become known after a very short time. It is not possible to perform a permanent modification in addition to the tests for career selection with the limited capacities available for such updates.

Requirements for the Diagnostic Process

Cognizance of these problems of aptitude diagnoses as well as the daily practice in the FAF provides a basis for the following requirements for future diagnostic work:

1. Improvement of the diagnoses is necessary; greater importance should be given to the differential diagnoses. A useful method should be found for solving the "bandwidth fidelity dilemma" so that, in spite of the use of mass testing, differential decisions are possible ("the right person in the right place"). This problem will become urgent for the FAF from about 1985 onward, when there will not be enough draftees available because of the rapid decline of the birthrate in the late 1960s.

2. Paper-and-pencil tests alone will not suffice in the future; with the improvement of diagnoses and the consideration of further aspects, skills, and experiences, it will be necessary to include new testing procedures and to test other psychological dimensions. Additionally, interests, motivations, and personality aspects should be tested, and perception and motor tests should be carried out to make more perfect diagnoses.

3. In addition to the test result--the score or ability parameter--other data should be included in the diagnostic process. Therefore, research programs concerning the testing process are necessary, including item solution time or time needed for solving a subtest, so that testing protocols (e.g., for counseling) can be produced.

4. Finally, mass testing makes it necessary to develop economy in the entire testing process and aptitude diagnosis. Scores and other computations should be carried out during the session, and results should be directly available at the end of testing. With these procedures and the proposed applications of items and subtests, it will be possible to save time and, moreover, to improve the diagnostic process.

Potential Solutions

Computerized testing will provide solutions to these problems in the following three areas:

Item production. Parts of item production can be performed by computer-assisted test construction (CATC). In a separate project, software was produced and implemented for item production and for individualization of tests, modifying the tests by computer. The first computer tests are in the empirical phase, and extensive results are expected in 1980.

Test data. For computation and interpretation of test data (selection and decision; "the diagnostic process") multi-faceted aids are possible: Simulations are being used in the FAF for computerized decision-making, and possibly the test results will be used to call up draftees.

Use of tests, the presentation of items, and scoring procedures. In addition to the presentation and computation of items using the classical concepts, there is a special case of test application--Computerized Adaptive Testing (CAT). Considerable savings and improvements of the aptitude diagnoses in the FAF are expected, especially from the adaptive methods and the new techniques of CAT.

Components of Computerized Testing

For the planning stage and implementation of computerized testing in the FAF a catalog was produced, containing the most important components of computerized testing and therewith also of CAT. These components, some of which will be empirically investigated by the FAF, include the following (the minimum requirements are preceded by an *):

Hardware

The requirement is defined to set up a test station, for example, for 50 testees carrying out diagnostic procedures of draftees. Many technical details (e.g., conception: connection to a large-size computer or stand-alone terminal station or a microprocessor for each testee; CPU and periphery, special screen and keyboards, and other facilities) are clarified and compared. Different products will be rated with regard to the requirements in the FAF.

1. *Requirement for flexibility of technology (e.g., extensions, innovations) and modular concept of hardware;

2. *Concept of the test station (connection to a large-size computer or stand-alone computer or microprocessor for each testee); system of minicomputers with foreground (input/output operations) and background (e.g., computations, estimations); multi-tasking, multi-processing;

3. *Central Processing Unit/Core Memory: construction, capacity/size, response/access/cycle time (e.g., station with 50 terminals testing draftees); byte or word, bit per word; accuracy/precision; floating point arithmetic (hardware or software, binary or decimal; number of bits for parameter estimations); *real-time execution, system-response time (processing an input immediately, without delay time);

4. *A printer for each testing station (production of testing protocols, plots); console display; possibility of storage, capacity of disks (magnetic disks or floppy disks); other storage on periphery; access mode and time; *archives/output of the raw data, compatibility (making copies to magnetic tapes, computations on an IBM large-size computer); processing the data on line/off line (among others, for the personnel division, using the test data in the data bases); connection to other computers in the FAF; definition of interfaces;

5. Equipment for a testing station for each testee (number of places connected to one processor); *special displays for presentation and processing the items; special keyboards (only digits and few buttons); display quality (sharp definition, contrast); graphic with 200,000 points, color equipment; use of video pictures; periphery, connection of further equipment/devices (tachistoscope, light pencil for figural tests or labyrinth items); usage of other apparatus or testing additional psychological dimensions with hardware or/and software (e.g., determination tool); controlling the testing process by acoustic stimulus, input of the answers using the terminal keyboard; employment of an A/D converter, making digitals using the physiological data or further testee data from other equipment;

6. *Infrastructure (e.g., power, power consumption, air conditioning); *mobility, possibility for transportation when testing draftees at different locations.

Test Applications/Concepts

The type of aptitude diagnoses to be taken over by a computer needs to be specified, for example, which psychological dimensions should be tested, which contents and methods should be used during the pilot projects (among others, the item-response time for ability estimation), and which further tasks (e.g., next item presentation and scoring) are possible with the test station, for example, computerized decision-making or counseling aspects.

1. *Flexibility for using different tests or methods; flexibility for time limits, sequence of subtests, power/speed tests, types of items, item material; flexibility for different data, changing the input of the test station (e.g., insertion of personal data or item-solution times); recording further psychological dimensions (perception, motor skills, concentration, coordination, fatigue, curves of learning, tracking); recording of interests, motivations, personality aspects; *possibility for different testing processes, omnibus procedures versus criterion-referenced measurement;

2. *Application of tests using the classical concept, presentation of conventional items by display (such as the present procedure for draftees); *jumping to different items, similar to the paper-and-pencil application (selection of different items by the testee, jumping forward and backward, as in a test book); *usage of sequential strategies based on subtests (screen and main test; indication of "critical

items" for the next subtest or for use in the interview); *processing the tests in groups of testees but continued application of individual tests/subsets of items;

3. Testing the pyramidal approach with the self-scoring aspect (in sensu Hornke, 1978); *application of tests with variable branching strategies, using different methods, different algorithms for parameter estimations, different scoring procedures, different criteria for cutoff; solving different methological problems using different estimation procedures (Bayesian, maximum likelihood, and so forth); inquiry of CPU/execution time using different methological approaches (different probabilistic models, various software);

4. Input of additional criterion data (e.g., age, date of final graduation from school), interests, special knowledge; recording the biographic data (using a questionnaire or free responses); *immediate computation of test data during the test process so that results are finished at the end of the session (i.e., scoring, norm values); interpretation of the test data, computerized diagnoses (classification with discriminant or cluster analyses); decision-making, placement recommendation for the draftees, taking into consideration the different requirements, priorities, or various criterion data of the armed forces; computerized personnel management (in contact with the data bases for the military personnel in the FAF); additional use of the test station for counseling aspects (e.g., possibilities of career, study at the universities of the FAF);

5. Possibility for giving feedback, processing several subtests; noting time limit if tests with time limit are in use (rest time per subtest, time used per item); *recording the item-solution time and processing the time as an additional ability estimator or for counseling; producing testing protocols with the response patterns (method for solving the subtest);

6. Possibility of computerized test construction; computation of follow-up analyses, validity approaches, and so forth.

Software

The system and the assembler programs monitoring the microcomputer, the possibilities for updating, the compatibility to an IBM large-size computer for follow-up analyses, and the real-time execution for presentation and computation of items should all be considered.

1. *Requirement for a modular system of software, implementation of new methods and testing procedures within a short time;

2. *Conversational/dialog program for processing the test sessions (selection of items, presentation, and computation; processing the item-solution time; possibly giving feedback); supervision of the test station (e.g., input/output, computations, interruptions, error han-

dling); monitoring the test process, operating log (e.g., internal statistics for usage of the subtests, items, error for handling, CPU time); *introduction for handling the CRT and the keyboard, processing of examples, operating the keyboard by various types of items; *check of the input for formal correctness (e.g., only one digit permissible or only a digit less than 5);

3. Requirement for programming the minicomputer by the user (e.g., the psychologist); using the higher software languages, such as FORTRAN or BASIC (interpreter or compiler); installation of a compiler for all stations or only for the development institution; usage of overlay techniques or virtual storage concepts, optimizing the core capacity; expense for programming, implementation of new tests, new methods, new software; support by utilities; improving the software and the assembler programs; updating the system of the minicomputer (e.g., presentation of the items, data management to archive the raw data, initial calculations); *storage of the data for follow-up analyses, transferring to a file of a large-size computer (calculations by SPSS or other software), development of software using a large-size computer via teleprocessing, simulating the minicomputer (e.g., conversational processing, compiler, assembler);

Organization and Usage

Checkpoints are the organization of the testing session during the entire selection process (with sport examination, medical check-up by physician, interview by psychologist, and so forth), operating the test station and the single screen/keyboard, handling for system trouble, and maintenance services.

1. *Requirement for simplicity of operations (nonspecialized operation of the testing station); *explanation for handling the CRT and keyboard for the testees (e.g., input, corrections, skipping forward and backward, giving assistance by a function HELP); monitoring the test process using the classical concept, i.e., for side-by-side terminals, parallel versions are presented;

2. Handling the test station for system trouble; restarting/restoring the system, rerunning the session, continuing with similar items (controlling the last transfer operation, the last processed item; successful processing of the last written operation; security of data (safe dump of the the raw data);

3. Breakdown time; maintenance services, agreement; spare parts; requirement for high readiness of operations;

4. Cost for purchase or lease, for maintenance and spare parts, and for operation (price-performance ratio);

5. Specific points of the firms (special features not described by the requirements above).

Planning and Procedure

Following the information and concept phase, in which information is collected and redefined for application of adaptive tests using a computer and for incorporation into the FAF, the first research programs are planned for examination and trial of the different contents, methods, and techniques; and the corresponding pilot projects are prepared. After checking computerized test applications in the FAF--their methods and techniques--the following parts and steps are designed:

1. Application of tests using the classical concept; presentation of conventional items by display; research program for the "psychology of computerized testing"; an experiment by Birke (1979) on the use of item-solution time as an additional ability estimator;

2. Testing the pyramidal approach with self-scoring (in sensu Hornke, 1977, 1978); and

3. Application of tests with variable branching strategies using different methods and approaches.

For these pilot projects item pools have been prepared and larger tests/subtests are presently in preparation. The extensive software should be produced in FORTRAN using the existing TSO connection to an IBM computer 370/168 and simulated corresponding test applications. Parallel to the planning of content and methods is the procuring of hardware, considering the components as previously designated in the catalog.

Since last year the FAF has had intensive contacts with the German firms Zak and Hogrefe, which--after many years of experience with the production of psychological-physiological tools--have offered microprocessor-based stand-alone computers for test application and for analyses of physiological data. Both firms are in the development phase, thus all offers have still not been realized (e.g., graphic equipment for 200,000 points, light pencil, use of video-tapes). Zak offers a modular system with 10 intelligent terminals, two floppy disk drives, and the central processor for one station; whereas Hogrefe offers a screen, a CPU, and a floppy disk for each testee.

The developments in the market are being observed and checked. Based on the requirements of the FAF, directions and concomitant requests are being formulated, and the German Ministry of Defense is providing a test station for the first pilot project for computerized test application.

Conclusion

The Psychological Services of the FAF is today at a starting-point of a new, rapid development of the testing process and aptitude diagnoses. At this time there is neither background experience nor a special approach to computerized testing in the FAF. Until now, problems of discussion and research designs have been oriented toward the practice in the FAF, derived from the everyday aptitude diagnoses requirements. I am certain, however, that in the coming years the traditional concept of testing by using paper and pencil will be eliminated.

REFERENCES

Birke, W. *Non-parametric measurement of intelligence on the basis of item-response-time: Outline of a model*. Unpublished manuscript, 1979.

Birke, W., & Wildgrube, W. *Die anwendbarkeit computerunterstützter adaptiver testmethoden im bereich der Bundeswehr* (Wehrpsychologische Mitteilungen 1/78). Bonn: Bundesministerium der Verteidigung Psychlogischer Dienst der Bundeswehr, 1978.

Buchtala, W. Entwicklung eines testinstrumentes für individualisierte testung durch adaptive parameterschätzung. In W.H. Tack, *Bericht über den 30. Kongress der Deutschen Gesellschaft für Psychologie in Regensburg* (Band 2). Göttingen: Hogrefe, 1977.

Fritscher, W., & Koch, E.W. *Elektronische datenverarbeitung und personalklassifikation im Psychologie dienst der Bundeswehr* (Wehrpsychologische Untersuchungen 3/75). Bonn: Bundesministerium der Verteidigung Psychologischer Dienst der Bundeswehr, 1975.

Hornke, L.F. Antwortabhängige testverfahren: Ein neuartiger ansatz psychologischen testens. *Diagnostica*, 1977, *23*, 1 - 14.

Hornke, L.F. Vergleich zweier adaptiv-antwortabhängiger testsrategien. *Diagnostica*, 1978, *24*, 103 - 112.

Steege, F.W. Personnel psychology in the Federal Armed Forces of Germany. In J.W. Miller (Ed.), *The 12th International Symposium on Applied Military Psychology* (C-26-27). London: Office of Naval Research, London Branch Office, 1976.

Weiss, D.J. (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Wildgrube, W. *Computergestützte hilfen bei der eignungsdiagnostik--problemaufriss* (Wehrpsychologische Mitteilungen 1/78). Bonn: Bundesministerium der verteidigung, Psychologischer Dienst der Bundeswehr, 1978.

Wildgrube, W. *Computergestütztes adaptives testen (CAT)--Neueste entwicklungen* (Arbeitsberichte Nr P-3-78). Bonn: Bundesministerium der Verteididung, Psychologischen Dienst der Bundeswehr, 1978.

SESSION 3:
ADAPTIVE MASTERY TESTING

SOME DECISION PROCEDURES FOR USE WITH TAILORED TESTING

MARK D. RECKASE
UNIVERSITY OF MISSOURI--
COLUMBIA

A MODEL FOR COMPUTERIZED ADAPTIVE TESTING RELATED TO INSTRUCTIONAL SITUATIONS

STANLEY J. KALISCH
EDUCATIONAL TESTING
SERVICE, ATLANTA

A COMPARISON OF ICC-BASED ADAPTIVE MASTERY TESTING AND THE WALDIAN PROBABILITY RATIO METHOD

G. GAGE KINGSBURY AND
DAVID J. WEISS
UNIVERSITY OF MINNESOTA

DISCUSSION

MELVIN NOVICK
UNIVERSITY OF IOWA

SOME DECISION PROCEDURES FOR USE WITH TAILORED TESTING

MARK D. RECKASE
UNIVERSITY OF MISSOURI

There are many applications of testing technology that require decisions to be made as to whether a person is above or below a criterion score. Criterion-referenced testing and its special case, mastery testing, are examples of such a decision. In the criterion-referenced testing application, it would be especially useful if decisions could be made quickly and conveniently for each student in an individualized instruction program. The recently developed technology of tailored/adaptive testing (Lord, 1970) has the potential to fulfill the requirements of such a testing system. However, there is no generally accepted procedure for making classification decisions using tailored testing, probably because these testing techniques are still relatively new. The few procedures that do exist are either based on randomly sampling items (Epstein, 1978; Sixtl, 1974), which does not take advantage of the power of tailored testing, or on heuristic techniques (Weiss, 1978), which do not have a sound theoretical base. The purpose of this paper is to present some decision procedures that operate sequentially and can easily be applied to tailored testing without loss of any of the elegance and mathematical sophistication of the examination procedures.

Tailored Testing Procedures

Numerous tailored (i.e., adaptive, response contingent, sequential) testing procedures now exist in the research literature, ranging from simple two-stage procedures (Betz & Weiss, 1973) to complex Bayesian procedures (Owen, 1969; see Weiss, 1974, for a good review of the tailored testing procedures that were developed prior to 1974.) Although many procedures exist, for the purposes of this paper only tailored testing procedures using item characteristic curve (ICC) theory and maximum likelihood ability estimation will be considered. It will also be assumed that the tests are administered to the examinees on a computer terminal and that the items are selected to maximize the value of the information function at the previous ability estimate. Despite the narrow definition of tailored testing used for this paper, the results should generalize to any procedure based upon ICC theory.

In applying the decision procedures discussed in this paper, two specific ICC models will be used: the 1- and 3-parameter logistic models. Although any other ICC model could just as easily have been used, these models were selected because of their frequent appearance in the research literature and because of the existence of readily available calibration programs (LOGIST, CALFIT) and tailored testing programs (Reckase, 1974).

Sequential Decision Procedures

A cursory review of the statistical literature indicates that much has been written about sequential estimation and classification procedures. Although somewhat more obscure than ANOVA and regression procedures, most intermediate level mathematical statistics books include at least one chapter on sequential analysis (for example, see Brunk, 1965, chap. 16). In an ongoing review of the extensive literature on this topic, it has been found that most procedures fall into one of three categories: 1) sequential probability ratio tests (SPRT; Wald, 1947), (2) Bayesian sequential procedures (e.g., DeGroot, 1970), and (3) curtailed single sampling plans (Dodge & Romig, 1929). Of these procedures, only the SPRT is narrowly specified--the other two refer to families of procedures rather than a single technique.

Although these statistical procedures are widely applied for quality control, little use has been made of them in the area of mental testing, probably because operable sequential testing procedures did not exist until recently. To date all references in the testing literature to sequential decisions have used the SPRT (Epstein, 1978; Reckase, 1978; Sixtl, 1974). The SPRT will therefore be described first, followed by the Bayesian procedures, since the curtailed sampling plans cannot readily be applied to the commonly used tailored testing procedures, they will not be discussed in this paper.

The Sequential Probability Ratio Test

The sequential probability ratio test (SPRT) was initially developed by Wald (1947) as a quality control device for use by the Armed Forces during World War II. In addition to Wald's (1947) excellent book on the subject, this procedure has been clearly described by Epstein (1978). It will, therefore, be only briefly described here in order to generalize the procedure so that it will more directly apply to tailored testing.

Application to Mastery Decisions

Wald originally developed the SPRT as a statistical test to decide which of two simple hypotheses is more correct. For example, it might be interesting to determine whether a student can answer correctly 60% or 80% of the items in an item pool. The basic philosophy behind the procedure used to decide between these two alternatives was to determine the likelihood of an observed response to an item under the two alternative hypotheses. If the likelihood were sufficiently larger for one hypothesis than the other, that hypothesis would be accepted. If the two likelihoods were similar, another observation would be taken. Wald (1947) has shown that one hypothesis will always be selected over another using a finite set of items.

To demonstrate this procedure, suppose an item is randomly selected from an item pool and administered to a student. If a correct response were obtained, the likelihood under H_1 (80% knowledge) would be .80, and the likelihood under H_0 (60% knowledge) would be .60. To evaluate these likelihoods, Wald takes the ratio of the two,

$$\frac{L(x = 1|H_1)}{L(x = 1|H_0)} = \frac{.80}{.60} \approx 1.67 . \qquad [1]$$

If the ratio is sufficiently large, H_1 is accepted; if it is sufficiently small, H_0 is accepted; and if it is near 1.0, another observation is taken. The values of this ratio that are considered sufficiently large or small depend upon what is considered acceptable for the two possible decision errors: (1) accepting H_1 when H_0 is true (α error) and (2) accepting H_0 when H_1 is true (β error).

Although Wald (1947) developed a procedure for determining the exact values of these decision points, the procedure is very complex and is seldom used. Instead, good approximations can be determined using the following formulas:

$$\text{lower decision point} = B = \frac{\beta}{1 - \alpha} \qquad [2]$$

$$\text{upper decision point} = A = \frac{1 - \beta}{\alpha} . \qquad [3]$$

Thus, if the likelihood ratio is less than or equal to B, H_0 is accepted with error probability approximately β. If the likelihood ratio is greater than or equal to A, H_1 is accepted with error probability approximately α. If the ratio is between B and A, another item should be randomly sampled and administered and the decision rule implemented again. If $\alpha = .05$ and $\beta = .10$, for example, the decision points would be at B = .105 and A = 18. Since the likelihood ratio (1.67) is between these two values, no decision would be made, and another item would be selected and administered.

Since the responses to the items follow a binomial distribution in this example, a general expression for the likelihood ratio can be developed for the administration of n items:

$$\frac{L(x_1, x_2, \ldots, x_n | H_1)}{L(x_1, x_2, \ldots, x_n | H_1)} = \frac{p_1^{\Sigma x_i}(1 - p_1)^{n - \Sigma x_i}}{p_0^{\Sigma x_i}(1 - p_0)^{n - \Sigma x_i}}$$

$$= \left(\frac{p_1}{p_0}\right)^{\Sigma x_i} \left(\frac{1 - p_1}{1 - p_0}\right)^{n - \Sigma x_i} , \qquad [4]$$

where

x_i is the score on item i (0 or 1),
p_1 is the proportion of items known by the student in the item pool under H_1, and
p_0 is the proportion known in the item pool under H_0.

If

$$\frac{L(x_1, \ldots, x_n | H_1)}{L(x_1, \ldots, x_n | H_0} \geq A, \text{ accept } H_1. \qquad [5]$$

If

$$\frac{L(x_1, \ldots, x_n | H_1)}{L(x_1, \ldots, x_n | H_0} \leq B, \text{ accept } H_0. \qquad [6]$$

Otherwise, continue administering items.

This procedure was originally developed to test simple hypotheses, but Wald (1947) has shown that the procedure operates in the same way for composite hypotheses. For example, suppose it is desirable to know whether a student knew more than some proportion, $\underline{p}_1$, of the items in an item pool. In order to use the SPRT to make this decision, a region for which it does not matter which decision is made must first be selected around $\underline{p}$, say, $\underline{p}_0 < \underline{p} < \underline{p}_1$. If $\underline{p}_0$ is close to $\underline{p}_1$, a very precise decision is required. If $\underline{p}_0$ and $\underline{p}_1$ define a wide indifference region around $\underline{p}$, a rather gross decision rule is all that is needed. The SPRT is then carried out in exactly the same fashion as above, using $\underline{p}_0$ and $\underline{p}_1$ as the values for hypotheses H_0 and H_1 ,respectively. When the decision points A and B are computed as above, the error rates, α and β, hold for true values of $\underline{p}$ at $\underline{p}_0$ and $\underline{p}_1$. For true values of $\underline{p}$ more extreme than $\underline{p}_0$ or $\underline{p}_1$, the error rates are lower.

Evaluating Outcomes

In order to evaluate the properties of the SPRT, two functions have been derived: the operating characteristic (OC) function and the average sample number (ASN) function. The OC function is defined as the probability of accepting hypothesis H_0 as a function of the true proportion of the item pool known by the student. Although the derivation of the OC function is somewhat complex, the function can be approximated by the following two formulas:

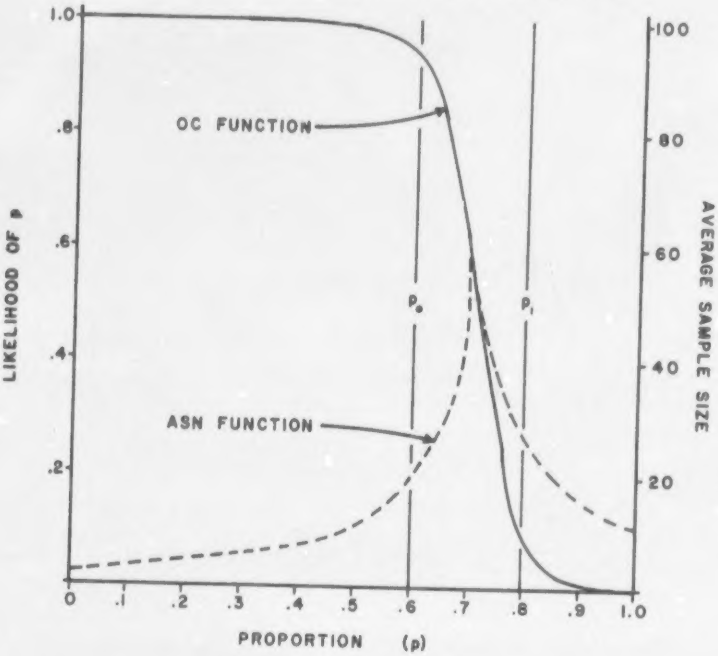
$$p = \frac{1 - \left(\frac{1 - p_1}{1 - p_0}\right)^h}{\left(\frac{p_1}{p_0}\right)^h - \left(\frac{1 - p_1}{1 - p_0}\right)^h} \qquad [7]$$

and

$$L(p) \simeq \frac{\left(\frac{1 - \beta}{\alpha}\right)^h - 1}{\left(\frac{1 - \beta}{\alpha}\right)^h - \left(\frac{\beta}{1 - \alpha}\right)^h} . \qquad [8]$$

These equations are used by substituting various arbitrary values of $\underline{h}$ and solving for $\underline{p}$ and L($\underline{p}$). L($\underline{p}$), the probability of accepting H_0, is then plotted against p to describe the OC function. Figure 1 shows an OC function for $\alpha = .05$, $\beta = .10$, $p_0 = .6$, and $p_1 = .8$. Note that at $p = p_0$ the height of the curve is equal to $1 - \alpha$, and at $p = p_1$, the height of the curve is equal to β. Note that the OC function is only dependent upon α, β, p_0, and p_1. Also, the steeper the curve, the more accurate the SPRT decision rule.

Figure 1
Example of the OC and ASN Functions



The ASN function is defined as the expected number of items required to make a decision at the various values of the true proportion of known items, $E(n|p)$. The formula for the ASN function for the binomial case described above is

$$E(n|p) = \frac{L(p)\ lnB + (1 - L(p))\ lnA}{p\ ln\left(\frac{p_1}{p_0}\right) + (1 - p)\ ln\left(\frac{1 - p_1}{1 - p_0}\right)}, \qquad [9]$$

where all of the symbols are as described above and the logarithms are to the base e. Figure 1 also shows the ASN function for the example presented above. Note that the ASN function is highest between the points p_0 and p_1 and that the closer together the values of p_0 and p_1 are, the higher the curve in that region. In general, the lower the ASN curve, the more efficient the decision rule.

Application to Tailored Testing

Although the SPRT as defined above is a valuable procedure for decision-making in many situations, it makes an implicit assumption that limits its usefulness for tailored testing. The model as presented assumes that the probability of a correct response is the same for all items in the pool. This assumption is reasonable if items are randomly selected and p is the proportion of the items that a student can answer correctly, but it is not reasonable if items are selected to maximize information at an ability level. Under the tailored testing model assumed by this paper, the probability of a correct response changes with each item, requiring a modification of the model.

Fortunately, a detailed analysis of Wald's (1947) work indicates that the sequential random sample assumption is not necessary for the application of the SPRT but is needed only for the derivation of the OC and ASN functions. The SPRT can then be directly applied to tailored testing, but the OC and ASN functions must be determined in a different manner. One approach to determining these functions will be presented later.

To demonstrate the application of the SPRT to tailored testing as defined by this paper, suppose that a tailored test is being used to determine whether a student has exceeded the criterion specified for a criterion-referenced test. Although the method for selecting this criterion is currently not well specified, assume that a value, θ_c, has been determined and that students above this value on the latent achievement scale pass the unit, while those below θ_c are given more instruction.

In order to use the SPRT, a region must be specified around θ_c for which it does not matter whether a pass or a fail decision is made. If high accuracy is desired for the decision rule, a narrow indifference region must be specified, but more items will be required to make the decision. As the region gets wider, the decision accuracy declines, but fewer items are required. Values of θ, θ_0, and θ_1 mark the boundaries of this indifference region ($\theta_0 < \theta_c < \theta_1$). Once these values have been selected, the likelihood ratio can be defined as

$$\frac{L(x_1,\ldots,x_n|\theta_1)}{L(x_1,\ldots,x_n|\theta_0)} = \frac{\prod_{i=1}^{n} P_i(\theta_1)^{x_i} Q_i(\theta_1)^{1-x_i}}{\prod_{i=1}^{n} P_i(\theta_0)^{x_i} Q_i(\theta_0)^{1-x_i}} \qquad [10]$$

where

$\underline{L}(\underline{x}_1,\ldots,\underline{x}_n|\theta_k)$, $\underline{k} = 0, 1$, is the likelihood of the student's response string of n items administered so far;

x_i is the 0, 1 score on item i;

$P_i(\theta_k)$ is the probability of a correct response to item i assuming ability θ_k determined from the appropriate ICC model; and

$Q_i(\theta_k) = 1 - P_i(\theta_k)$.

If the 1-parameter logistic model is used as a basis for the tailored testing procedure, Equation 10 becomes

$$\frac{L(x_1, \ldots, x_n | \theta_1)}{L(x_1, \ldots, x_n | \theta_0)} = \frac{\prod_{i=1}^{n} \dfrac{e^{x_i(\theta_1 - b_i)}}{1 + e^{(\theta_1 - b_i)}}}{\prod_{i=1}^{n} \dfrac{e^{x_i(\theta_0 - b_i}}{1 + e^{(\theta_0 - b_i)}}} , \qquad [11]$$

where b_i is the difficulty parameter for item i. Equation 11 can be simplified to

$$\frac{L(x_1, \ldots, x_n | \theta_1)}{L(x_1, \ldots, x_n | \theta_0)} = e^{\sum_{i=1}^{n} x_i(\theta_1 - \theta_0)} \prod_{i=1}^{n} \frac{1 + e^{(\theta_0 - b_i)}}{1 + e^{(\theta_1 - b_i)}} \qquad [12]$$

The values of this likelihood ratio can then be used to test whether the student is above or below θ_c using the same method presented earlier. If the ratio is greater than $A = \frac{(1 - \beta)}{\alpha}$. the student is classified as being above θ_c; if it is below $B = \frac{\beta}{(1 - \alpha)}$, the student is classified below the criterion; otherwise, another item is administered. If the 3-parameter logistic model is the basis for the tailored testing procedure, the SPRT procedure is applied in exactly the same manner as above, except that

$$P_i(\theta_k) = c_i + (1 - c_i) \frac{e^{Da_i(\theta_k - b_i)}}{1 + e^{Da_i(\theta_k - b_i)}} \qquad [13]$$

is used in Equation 10 instead of the simple logistic form.

The evaluation of the OC and ASN functions cannot be performed as easily as for the simple binomial model due to the presence of the item parameters in the formula for computing the probability of a correct response. Since the item parameters for the next item to be administered are dependent on the item pool used and on the responses to the previous items, the derivation of these functions depends on a complex string of conditional expectations. The conditional probabilities involved make the derivation of these functions, for all practical purposes, impossible. Therefore, the OC and ASN functions can only be approximated using simulation techniques, but these approximations should be adequate for most purposes. Some OC and ASN functions for tailored tests based on the 1- and 3-parameter logistic models will be presented later in this paper. Note, however, that although the full OC function cannot be derived, the value of the function is equal to $1 - \alpha$ at θ_0 and to β at θ_1, assuming that the item parameters are known. In reality, these two points are not known either, since in all cases except simulations the item parameters are only estimated.

Bayesian Sequential Decision Procedure

The Bayesian decision procedure is an alternative to the SPRT for deciding whether or not a student has exceeded the criterion, θ_c. Although this procedure is much more complicated than the SPRT, it has the capability of using additional information in making the decision. This added information may improve the decision process.

Basic Concepts

Initially, it is assumed that a population of students exists such that each student has some definable achievement level, θ. Individual achievement levels are labeled θ_i. Each person is to be tested and a decision is to be made concerning placement above or below the criterion. The decision to place above the criterion score is labeled d_1; and the decision to place below the criterion score, d_2.

In order to decide upon a decision rule using Bayesian methodology, three pieces of information are required in advance. These are (1) a prior distribution of θ, (2) a loss function relating the achievement levels to the decisions, and (3) the cost of each observation. Using these three types of information, a decision rule (technique for selecting a decision) and a stopping rule (technique for deciding when a decision should be made) can be determined.

The basic concept used in choosing a decision rule is the concept of risk. Risk is defined as the expected loss, given a decision. Obviously, the decision that minimizes the risk is the desired one. When a Bayesian prior is used, this minimum risk is called the Bayes risk.

The stopping rule used with the Bayesian sequential decision procedure is also based upon the Bayes risk concept. If the expected risk after taking another observation plus the cost of the observation is less than the risk before the observation is taken, the sampling should go on. However, if the expected risk plus the cost of a new observation is greater than the risk without the observation, then sampling should cease. In some cases, it is best not to take any observations at all, because the expected risk plus the cost of an observation is greater than the initial risk of a guess based on the prior distribution of achievement.

Based on this framework, theorems have been proven showing that an optimal procedure exists and that the optimal procedure will reach a decision after some finite number of observations (DeGroot, 1977). If the risk decreases with each observation, the procedure is called a regular sequential decision procedure. Only regular procedures will be considered here, since it is assumed that each item administered yields some positive information rather than providing some misinformation.

Simplified Example

Although this example is not realistic, it demonstrates the basic concepts without requiring complicated mathematical expressions. The extension of the procedure to realistic situations is direct, but the mathematics is cumbersome. Suppose that two types of individuals exist in the population of interest, those with $\theta_i = -.8$ and those with $\theta_i = +.8$ on a latent achievement dimension. A tailored test is to be used to classify the individuals into two groups--those above the criterion score 0.0 and those below. Thus, two decisions are possible: (1) classify as $\underline{d}_1$ those above the criterion and (2) classify as $\underline{d}_2$ those below the criterion.

If persons with ability -.8 are classified above the criterion, a loss of 25 is incurred in each case. If they are classified below the criterion, there is no loss. If persons with ability +.8 are classified above the criterion, there is no loss, whereas a loss of 15 is incurred for each person classified below the criterion. This loss function is summarized in Table 1; it should be noted that these loss function values are totally arbitrary.

Table 1
Loss Function

Ability (θ_i)	Decision $\underline{d}_1$	 $\underline{d}_2$
+.8	0	15
-.8	25	0

Suppose that the prior belief that a randomly selected person has ability +.8 is .6 and the prior belief that he/she has ability -.8 is .4. Then, the first step in using a Bayesian sequential decision process is to determine the risk associated with $\underline{d}_1$ and $\underline{d}_2$ when no observations are taken. The expected loss (risk) if decision $\underline{d}_1$ is made is

$$\begin{aligned} E(\text{loss}|d_1) &= P(\theta_1)\ell(d_1|\theta_1) + P(\theta_2)\ell(d_1|\theta_2) \\ &= .4 \times 25 + .6 \times 0 \\ &= 10, \end{aligned} \quad [14]$$

where $P(\theta_i)$ is the prior probability of θ_i and $\ell(\underline{d}_j|\theta_i)$ is the loss from making decision $\underline{d}_j$ when θ_i is true. The expected loss (risk) if $\underline{d}_2$ is made is

$$\begin{aligned} E(\text{loss}|d_2) &= P(\theta_1)\ell(d_2|\theta_1) + P(\theta_2)\ell(d_2|\theta_2) \\ &= .4 \times 0 + .6 \times 15 \\ &= 9. \end{aligned} \quad [15]$$

Thus, the Bayes decision when no observation is taken is $\underline{d}_2$, and the Bayes risk is 9. The decision $\underline{d}_2$ is obviously chosen because it has the lower risk.

Although the proper decision has been determined for the case when no observations have been taken, it has not been determined whether or not an observation should be taken. To do that, the expected risk after one observation plus cost must be compared to the Bayes risk without an observation. Determining the expected risk after an observation requires several steps, the first of which is determining the posterior distribution of ability after an observation.

Suppose that an item of 0.0 difficulty is administered to a person with ability +.8 or −.8. Depending upon whether the response is correct or incorrect, a Bayesian posterior can be determined using Bayes theorem

$$P(\theta_i|x) = \frac{P(x|\theta_i)\ P(\theta_i)}{\sum_{i=1}^{2} P(x|\theta_i)\ P(\theta_i)} \quad . \qquad [16]$$

If a correct response to the item is obtained, the posterior probability of a +.8 ability is given by

$$P(.8|x = 1) = \frac{P(1|.8)P(.8)}{P(1|.8)P(.8) + P(1|-.8)P(-.8)} \quad . \qquad [17]$$

The probabilities of an ability of +.8 or −.8 were given in the prior distribution as .6 and .4, respectively. The probability of a correct response, given the known ability, can be determined from the appropriate ICC model. For example, using the 1-parameter logistic model,

$$P(1|.8) = \frac{e^{(.8 - 0)}}{1 + e^{(.8 - 0)}} = .69 \ , \qquad [18]$$

where $P(1|-.8) = .31$. The posterior probability of +.8 is then $P(.8|1) = .77$. Similarly, the posterior probability of −.8 is $P(-.8|1) = .23$. The posterior probability of the +.8 and −.8 abilities, given an incorrect response, can likewise be determined using Equation 16. The posterior probabilities, given an incorrect response, are $P(.8|0) = .37$ and $P(-.8|0) = .63$.

The next step is to determine the risk using the posterior distributions just computed. If a correct response is obtained, the expected loss for $\underline{d}_1$ is .23 × 25 + .77 × 0 = 5.75. The expected loss for $\underline{d}_2$ is .77 × 15 + .23 × 0 = 11.55. Thus, if a correct response is obtained, the Bayes decision is $\underline{d}_1$ with a Bayes risk of 5.75. If an incorrect response is obtained, the expected loss for $\underline{d}_1$ is .63 × 25 + .37 × 0 = 15.75, while the expected loss for $\underline{d}_2$ is .37 × 15 + .63 × 0 = 5.55. Thus, after an incorrect response, $\underline{d}_2$ is the Bayes decision with a Bayes risk of 5.55.

Since it is not known whether a correct or incorrect response will be given, the expected risk must be computed regardless of the response. To compute the overall expected risk, the probability of a correct and an incorrect response is needed. The probability can be obtained using the following formula:

$$P(1) = P(1|.8)P(.8) + P(1|-.8)P(-.8) \qquad [19]$$

$$= .69 \times .6 + .31 \times 4$$

$$= .538$$

$$P(0) = 1 - P(1) = .462 .$$

The expected risk after a response can now be determined from

$$E(\text{risk}|\text{response}) = E(\text{loss}|1)P(1) + E(\text{loss}|0)P(0) \qquad [20]$$

$$= 5.75 \times .538 + 5.55 \times .462$$

$$= 5.66 .$$

At this point, whether or not another observation should be taken can be determined. If the expected loss after an observation plus cost is greater than the risk before an observation, then administration of items should cease. If the risk before an observation is taken is greater, then another item should be administered. In the example given here, assume the cost of a response is 1 unit. The expected loss after a response plus cost is then 5.66 + 1 = 6.66. Since the Bayes risk with no items administered was 9, another item should be administered. Depending on the response to the item, decision $\underline{d}_1$ or $\underline{d}_2$ could be selected. After the item is administered, the appropriate posterior becomes the new prior and the process continues as above. A flowchart of the entire decision process is presented in Figure 2.

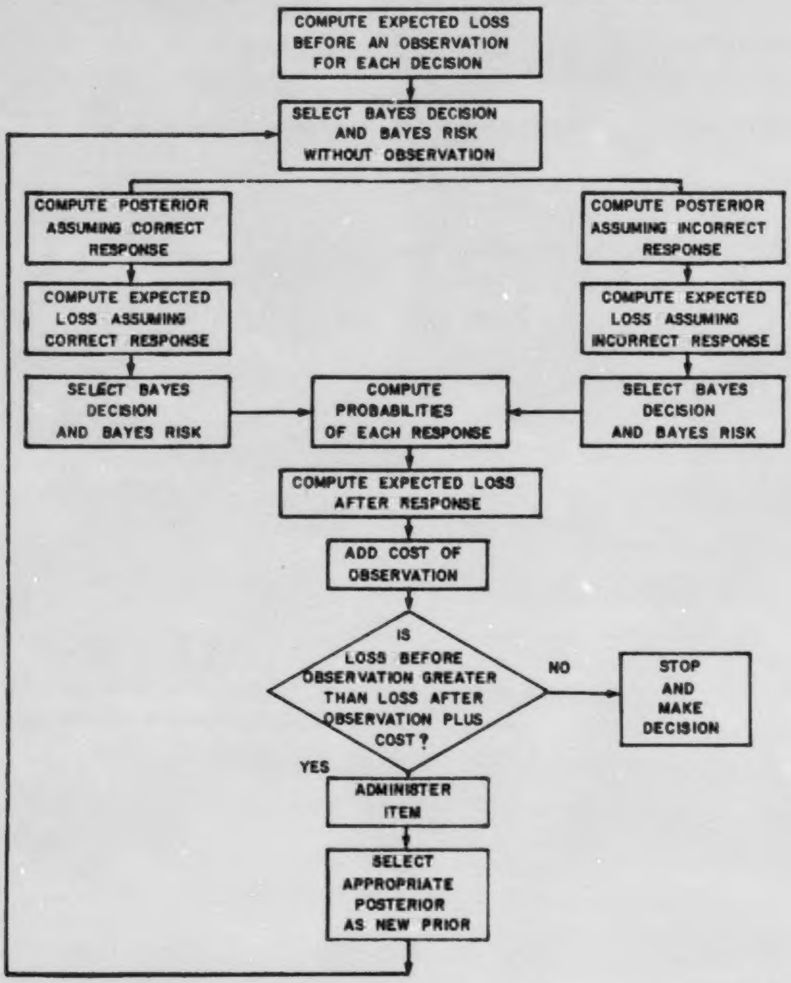
Limitations

Although there are many positive factors in the use of the Bayesian procedure, the very information that makes the control of the testing situation more precise also makes it difficult to implement initially. For example, specifying reasonable loss functions on the same metric as the cost of an observation is difficult for most educational applications. What is the cost of misclassifying persons below the criterion's score when they really should be classified above it? Some attempts have been made by this author to specify loss functions for tailored testing applications, but no satisfactory results have been obtained so far.

A second difficulty in the application of this procedure is in specifying the prior distribution of achievement for a group. This is not as serious a problem as determining loss functions, since performance data are usually available from previous groups. Of course, the more accurate the prior distribution, the more accurate the decision based on the procedure.

It should be realized that the procedure presented here is a simplification of a procedure that would be used for actual tailored testing applications. Achievement levels are usually continuous rather than discrete, as presented here; and the loss due to an incorrect decision is a function of the person's distance from the criterion score rather than a constant value. The procedure can also be modified by changing the cost of observations with increasing test length to allow for fatigue effects. Unfortunately, the Bayesian decision pro-

Figure 2
Flowchart of Bayesian Decision Process


COMPUTE EXPECTED LOSS BEFORE AN OBSERVATION FOR EACH DECISION
SELECT BAYES DECISION AND BAYES RISK WITHOUT OBSERVATION
COMPUTE POSTERIOR ASSUMING CORRECT RESPONSE
COMPUTE POSTERIOR ASSUMING INCORRECT RESPONSE
COMPUTE EXPECTED LOSS ASSUMING CORRECT RESPONSE
COMPUTE EXPECTED LOSS ASSUMING INCORRECT RESPONSE
SELECT BAYES DECISION AND BAYES RISK
COMPUTE PROBABILITIES OF EACH RESPONSE
SELECT BAYES DECISION AND BAYES RISK
COMPUTE EXPECTED LOSS AFTER RESPONSE
ADD COST OF OBSERVATION
IS LOSS BEFORE OBSERVATION GREATER THAN LOSS AFTER OBSERVATION PLUS COST?
NO
STOP AND MAKE DECISION
YES
ADMINISTER ITEM
SELECT APPROPRIATE POSTERIOR AS NEW PRIOR

cedure as described here has not yet been implemented in conjunction with an operational tailored testing procedure. Plans are being developed, however, to evaluate an operational version at the Tailored Testing Research Laboratory at the University of Missouri.

Research Design

The purposes of this research were (1) to obtain information on how the SPRT procedure functioned when items were not randomly sampled from the item pool; (2) to gain experience in selecting the bounds of the indifference region, θ_0 and θ_1; and (3) to obtain information on the effects of guessing on the accuracy of classification when the 1-parameter logistic model was used.

Tailored Testing Procedure

To determine the effects of these variables, the computation of the SPRT was programmed into both the 1- and 3-parameter logistic tailored testing procedures that were operational at the University of Missouri-Columbia. Since these procedures have been described in detail previously (Koch & Reckase, 1978), they will be merely summarized here. The programs implementing both models used a fixed stepsize method for branching through an item pool until both a correct and an incorrect response had been given. After that point, all ability estimates were obtained using an empirical maximum likelihood estimation procedure. Items were selected for both models to maximize the item information at the previous ability estimate.

To evaluate the decision-making power of the SPRT, subjects with known ability were needed. Therefore, a simulation routine was built into the tailored testing program in place of the responding live examinee. At the beginning of each simulation run, the true ability of the simulated examinee was input into the program. This value was used to determine the true probability of a correct response to the administered items based on the model used (1- or 3-parameter logistic) and the estimated item parameters. A number was then randomly selected from a uniform distribution in the range from 0 to 1. If the randomly selected number was less than or equal to the probability of a correct response, the item was scored as correct. If the randomly selected number was greater than the probability of a correct response, the item was scored as incorrect. This procedure continued for each item in the tailored test.

Tailored tests were simulated 25 times at each true ability using different seed numbers for the random number generator. True abilities from -3 to +3 at .25 intervals were used for both the 1- and 3-parameter models to evaluate the performance of the SPRT. In addition, simulations were run on a composite procedure in which tailored test procedure and the probability ratio calculations (Equation 11) were based on the 1-parameter model, but the item responses were determined by using the 3-parameter model. This was done to determine the effects of guessing on correct classification using the 1-parameter logistic model.

Criterion Values

In computing the probability ratios, three sets of limits of the indifference regions were used: $\pm.3$, $\pm.8$, ±1. A criterion of $\theta_c = 0$ was assumed in all cases. The ratios were computed after each item was administered, and the results were compared to an A value of 45 and a B value of .102. These were determined based on $\alpha = .02$ and $\beta = .10$. A classification was made the first time these limits were exceeded. If the limits were not exceeded before 20 items had been administered (an arbitrary upper limit on test length), the values above 1.0 were classified as above θ_c and the values below 1.0 were classified as below θ_c. This is called a truncated SPRT. At each true ability used for the simulation, the proportion of the 25 administrations classified below θ_c and the average number of items administered were computed. Plots of these values against the true abilities approximate the OC and ASN functions, respectively.

These plots were made for each combination of indifference region and tailored testing method, yielding nine plots of the OC and ASN functions.

Item Pools

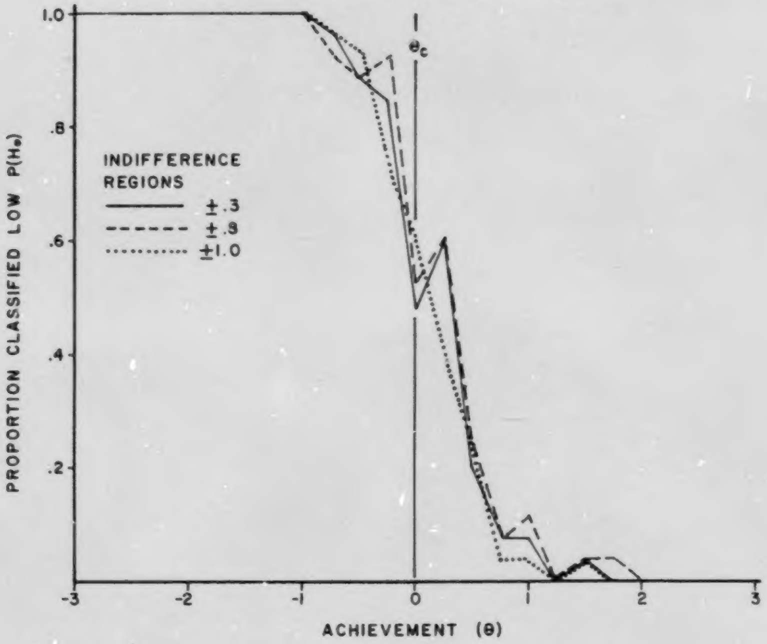
Two different item pools were used for this study. For the analyses using just the 1-parameter or the 3-parameter model, an existing pool of 72 vocabulary items were used. This item pool had an approximately normal distribution of difficulty parameters. For the 1-parameter tailored test using 3-parameter responses, an item pool with 181 items, rectangularly distributed between -3 and +3 in difficulty was used. These simulated items had constant discrimination parameters of .588 (this value yields a 1.0 when multiplied by D = 1.7) and a pseudo-guessing parameter of .12. This simulated item pool was selected over the real vocabulary pool to have better control over the guessing parameters. The 1-parameter procedure used only the b-values from the pool.

Results

1-Parameter Model

Figure 3 shows the OC functions for the 1-parameter logistic model based on the vocabulary item pool. The figure shows three graphs, one for each of the ±.3, ±.8, and ±1 indifference regions. Note that the curves are similar regardless of the indifference region. The data indicate that in all three cases the classification accuracy was nearly the same.

Figure 3
One-Parameter OC Functions
for Three Indifference Regions



The values of the curves at the limits of the indifference region give further evaluative information. At the lower point the OC function should pass through 1 - α. At the -.3 value the curve is in fact .85 when it should be .98, showing the degrading effects of restrictive stopping rules used by the tailored testing procedure. At the -.8 and -1 points for the corresponding curves, the results are about as expected, being .94 and 1.00 rather than .98.

At the upper limit of the indifference region, the OC function should have a value of .1. For the +.3 case it is in fact .5 rather than .1, again showing the effects of truncating the procedure. At the values of +.8 and +1 the values of the OC function were near or better than what they should have been, based on the theoretically expected results.

The ASN functions for the 1-parameter model are given in Figure 4. The curves plotted correspond to the ASN functions, using indifference regions for ±.3, ±.8, and ±1. It can immediately be seen that there was a substantial difference in the average number of items needed to reach a decision, with the greatest number required when the indifference region was narrowest. It can also be seen that the largest expected number of items was near the criterion score 0.0 and that the average number dropped off at the extreme abilities. The slight lack of symmetry in the curves is due to the fact that α was not equal to β. For abilities beyond ±1, an average of only about 3 to 5 items was needed for classification for the wider regions, but 6 to 11 items were needed for the ±.3 indifference region. Note that the ±.3 curve approached the arbitrary 20-item limit for the tailored tests.

Figure 4
One-Parameter ASN Functions
for Three Indifference Regions

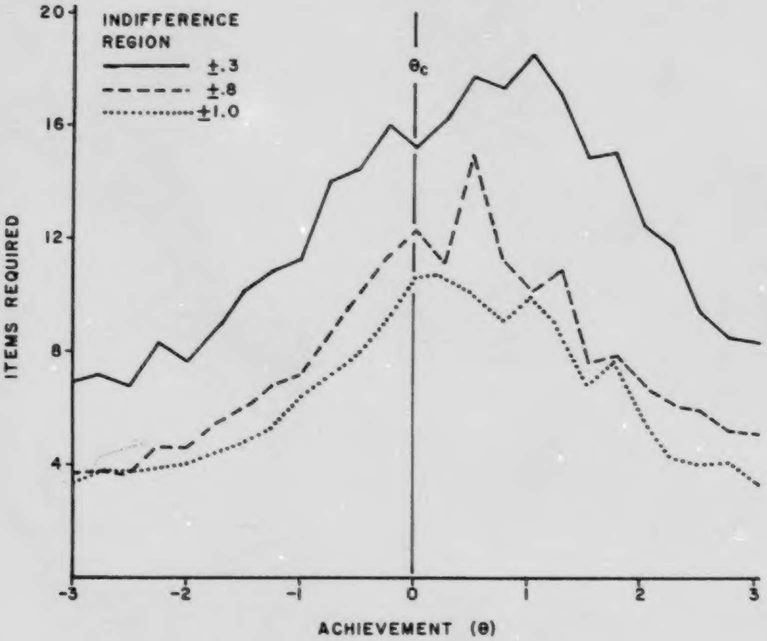
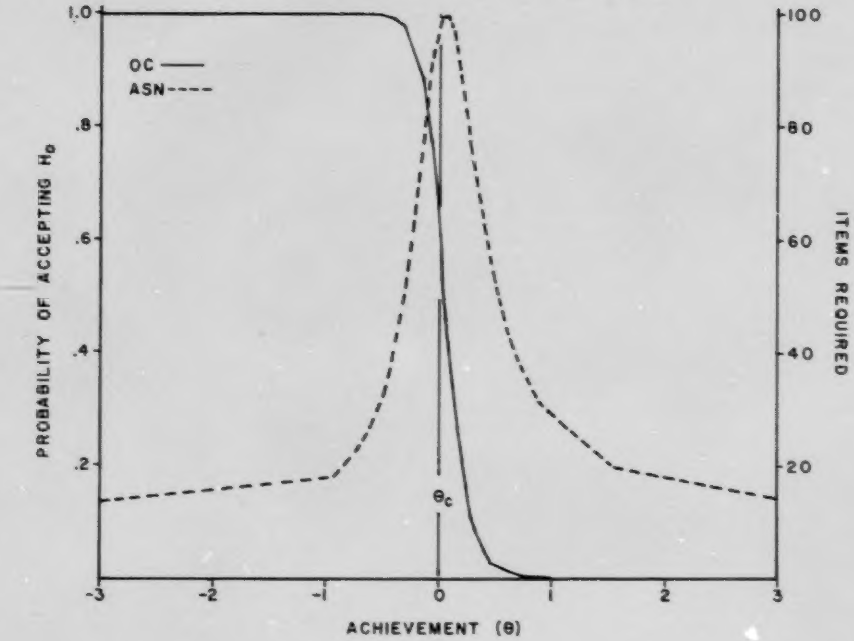


Figure 5 shows, for comparison purposes, the theoretical curves for the ASN and OC functions based on the ±.3 indifference region. An infinite number of items with difficulty 0.0 was assumed for the theoretical functions, and the tests were assumed to have no upper limit on the number of items administered. A comparison of Figures 3 and 4 with Figure 5 shows that the OC curve for the theoretical function is steeper at the cutting point than the simulated curves, and that the ASN function is substantially higher. The difference in the theoretical and simulated OC curves shows the effect of the 20-item stopping rule and the selection of items of differing difficulty.

Figure 5
Theoretical OC and ASN Functions

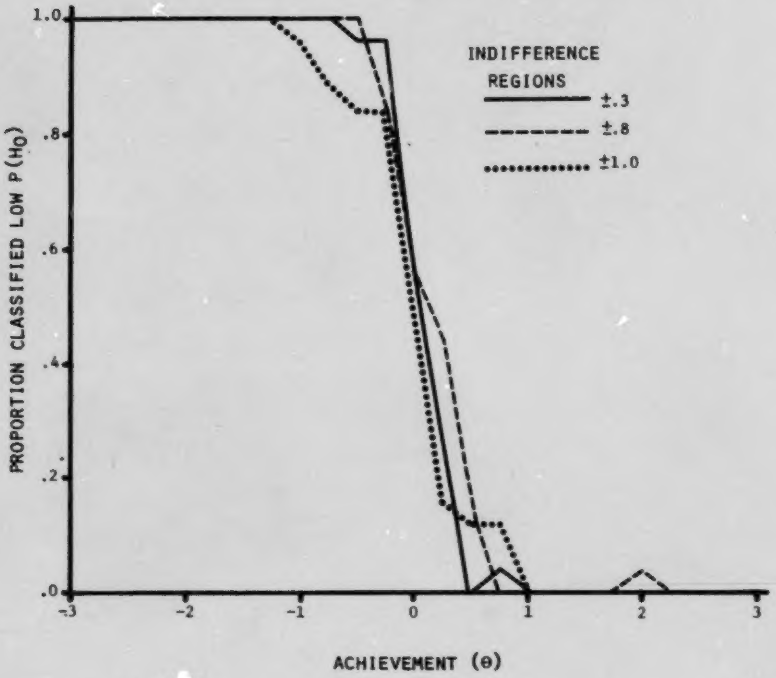


3-Parameter Model

The results of the simulation of the 3-parameter logistic tailored test are given in Figures 6 and 7. Figure 6 presents the OC functions for the 3-parameter model, again using the indifference regions of ±.3, ±.8, and ±1. Notice that as with the 1-parameter model, the OC curves are fairly similar for the three indifference regions throughout most of the range of ability. However, there are discrepancies for the ±1 indifference range curve near the +1 and -1 points, indicating a decline in decision precision for that region. At the -.3 value for the ±.3 indifference range, the value of the curve is .96, fairly close to the .98 theoretical value. At the upper end (+.3), however, the value is .2 instead of the .1 value that it should be. This may show the effects of guessing on the decision process. The ±.8 and ±1 indifference regions again yield better error probabilities than would be expected from the theory.

The ASN function for the 3-parameter model (Figure 7) also shows similar

Figure 6
Three Parameter OC Functions
for Three Indifference Regions



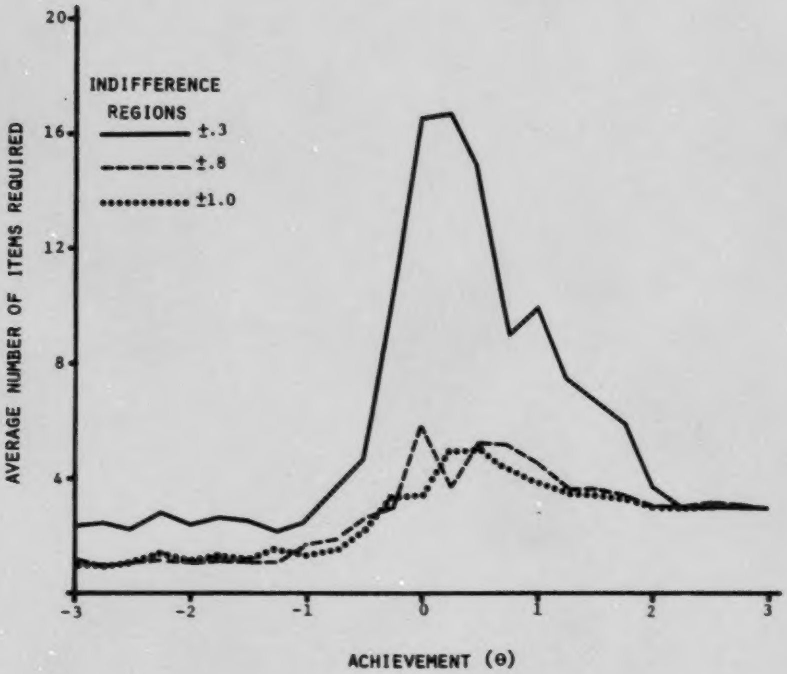
results to those obtained from the 1-parameter model. The +.3 indifference region required the greatest number of items, while +.8 and +1.0 required about the same number. As before, the largest number was required near the criterion score. However, with the 3-parameter model far fewer items, on the average, were required to make a decision than for the 1-parameter model. Of special note is the ASN value of about 1.0 in the -1 to -3 range on the ability scale. Decisions seem to be possible with very few items in that range.

Because of the guessing component of the 3-parameter logistic model, the ASN function tended to yield more asymmetric results than the 1-parameter model. More items were required when classifying high than when classifying low to compensate for the nonzero probability of a correct response. Also, the ASN curve for the +.3 indifference region was much more peaked than its 1-parameter counterpart. If the simulated curves for the 3-parameter model are compared to the theoretical curves presented in Figure 5, the OC functions can be seen to match the theoretical functions fairly closely, while the ASN functions show that substantially fewer items were required. Over much of the ability range, as many as 10 times more items were specified by the theoretical ASN curve when unlimited identical items were assumed. However, it should be noted that the theoretical curves are based on the 1-parameter model.

Effect of Guessing on the 1-Parameter Model

Figure 8 shows the OC functions for the 1-parameter model when the 3-param-

Figure 7
Three Parameter ASN Functions
for Three Indifference Regions



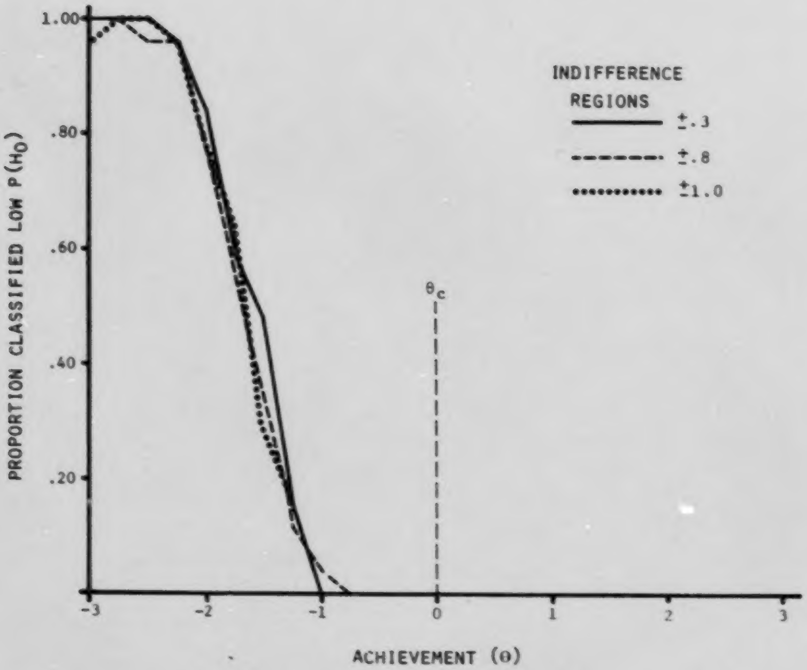
eter model was used to determine the responses. The figure shows three graphs, one for each of the ±.3, ±.8, and ±1 indifference regions. Note that the curves are fairly similar regardless of the indifference region but that they are shifted substantially to the left compared to the previous OC curves. This indicates that the probability of classifying a person below θ_c has dropped off substantially until an ability of about -2 has been reached. In other words, it is much easier to be classified above the criterion score with this procedure than when guessing does not enter into the decision. Instead of being at zero, the effective criterion has been shifted down to -1.5. Clearly, the values of the OC function at the limits of the indifference region are entirely different from the theoretical values.

The ASN functions for the three indifference regions—±.3, ±.8, and ±1—are shown in Figure 9. The difference between these graphs and those presented in Figure 4 are that the curves are higher (more items were required) and the highest point of the curve is shifted to the steepest part of the OC curve. The relationship between the height of the ASN function and the width of the indifference region still holds; however, as the region gets wider, the average number of items decreases.

Summary and Conclusions

The purpose of this paper has been to describe two procedures for making

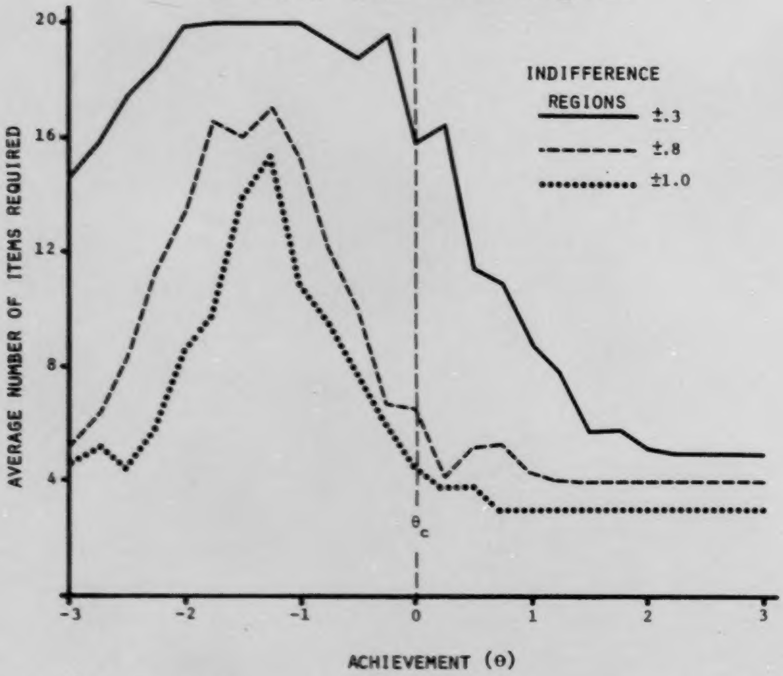
Figure 8
Composite OC Functions
for Three Indifference Regions



binary classification decisions using tailored testing--the sequential probability ratio test (SPRT) and a Bayesian decision procedure--and to present some simulation data showing the characteristics of the operation of the SPRT for two ICC models. The first procedure described, the SPRT, was developed by Wald for quality control work. It has not been widely applied for testing applications because the assumption of an equal probability of a correct response was made to facilitate the derivation of the operating characteristic (OC) and average sample number (ASN) functions. Since this assumption can only be met for testing applications by randomly sampling items for administration, the procedure has not been used with tailored testing. In this paper the probability of a correct response was allowed to vary from item to item, although it made the derivation of the OC and ASN functions impossible. Simulation procedures were then used to estimate these functions.

The SPRT procedure described is operational at the Tailored Testing Research Laboratory of the University of Missouri-Columbia in two forms: a live tailored testing procedure and a simulated procedure. The results of the application of the simulation procedure to three studies were described in this paper. The first study estimated the OC and ASN functions for a 1-parameter logistic based tailored testing procedure in which the size of the indifference region around the criteron score was varied. The results of the study showed that the average number of items needed for classification was quite low when the true ability of a simulated person was not too close to the criterion score

Figure 9
Composite ASN Functions
for Three Indifference Regions



and that the width of the indifference region did not greatly affect the OC function. The width of the indifference region did have a substantial effect on the ASN function. The accuracy of classification of the simulated tailored test was not quite as good as administering a large number of items with difficulty values equal to the criterion score. This result was explained by the arbitrary 20-item limit imposed on the tailored test and by the variation in the difficulty parameters of the items administered.

The second study estimated the OC and ASN functions for a 3-parameter logistic tailored testing procedure, also varying the size of the indifference region. The results were similar to those for the 1-parameter model, but even fewer items were generally needed for classification. The results of these first two studies both indicated that the SPRT could be successfully applied to tailored testing.

The third simulation study estimated the OC and ASN functions for the 1-parameter model when guessing was allowed to enter into the responses to the items administered. The results showed that, in effect, guessing lowered the criterion score, making it easier to classify an examinee above the criterion and raising the average number of items needed for classification. This spurious shift in the criterion greatly increased the error rates in classification. The effect was strong enough to preclude the use of the 1-parameter model for classification decisions when guessing is a factor.

The second decision procedure described in this paper allows the use of a greater amount of information in making a decision than the SPRT. The Bayesian procedure includes a prior distribution of student achievement, a loss function for incorrect decisions, and the cost of observations in the development of the decision rule. The basic philosophy of this procedure is to administer items until the expected loss incurred in making a decision is less than the expected loss after the next item is administered plus the cost of administration. At that point a decision is made that minimizes the expected loss. The Bayesian procedure is described in detail, and a simple example is given of its use. The Bayesian procedure is not yet operational for making decisions under tailored testing because appropriate loss functions for educational decisions have not been determined. However, simulation studies of the procedure will commence in the near future.

Both of the decision procedures described in this paper show promise for use in tailored testing. Both also require substantial research effort before they can be applied with confidence. It is hoped that this paper will help to stimulate that research.

REFERENCES

Betz, N. E., & Weiss, D. J. *An empirical study of computer-administered two-stage ability testing* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1973. (NTIS No. AD 768993)

Brunk, H. D. *An introduction to mathematical statistics* (2nd ed.). New York: Blaisdell, 1965.

DeGroot, M. *Optimal statistical decisions*. New York: McGraw-Hill 1970.

Dodge, H. F., & Romig, H. G. A method of sampling inspection. *Bell System Technical Journal*, 1929, *8*, 613-631.

Epstein, K. Applications of sequential testing procedures to performance testing. *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978.

Koch, W. R., & Reckase, M. D. *A live tailored testing comparison study of the one- and three-parameter logistic models* (Research Report 78-1). Columbia: University of Missouri, Tailored Testing Research Laboratory, June 1978.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance*. New York: Harper & Row, 1970.

Owen, R. J. *A Bayesian approach to tailored testing* (Research Bulletin RB-69-92). Princeton, NJ: Educational Testing Service, 1969.

Reckase, M. D. A generalization of sequential analysis to decision making with tailored testing. Paper presented at the meeting of the Military Testing Association, Oklahoma City, November 1978.

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods and Instrumentation, 1974, 6, 208-212.

Sixtl, F. Statistical foundations for a fully automated examiner. Zeitshrift fur Entwicklungspsychologie und Padagogische Psychologie, 1974, 6, 28-38.

Wald, A. Sequential analysis. New York: Wiley, 1947.

Weiss, D. J. Presentation at the ONR Contractors' meeting. Columbia: University of Missouri, September 1978.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1974. (NTIS No. AD A004270)

ACKNOWLEDGMENT

This research was supported by Contract No. N00014-77-C0097 from the Personnel and Training Research Programs of the Office of Naval Research.

A MODEL FOR COMPUTERIZED ADAPTIVE TESTING RELATED TO INSTRUCTIONAL SITUATIONS

STANLEY J. KALISCH
EDUCATIONAL TESTING SERVICE, ATLANTA

The present study involved the formulation and evaluation by computer simulation of a model for computer-based adaptive testing related to instructional or training situations. Specifically, the model adresses tests composed of items corresponding to hierarchically related instructional objectives. The purpose of the endeavor was to formulate and to analyze a model that would reduce testing time without compromising the necessary level of accuracy in decisions regarding the mastery or nonmastery of objectives.

The adaptive testing model developed in this study combines the models of Ferguson (1969, 1970) and Kalisch (1974a, 1974b). Ferguson's procedure employs the Wald probability ratio test (Wald, 1947, 1973) to determine mastery/nonmastery of hierarchically related objectives. Kalisch's procedure employs a process that predicts item responses based upon prior examinees' data. For the present study a combination of obtained and predicted item responses was used with the Wald binomial probability ratio test and hierarchical configurations of objectives to ascertain each examinee's mastery/nonmastery of objectives.

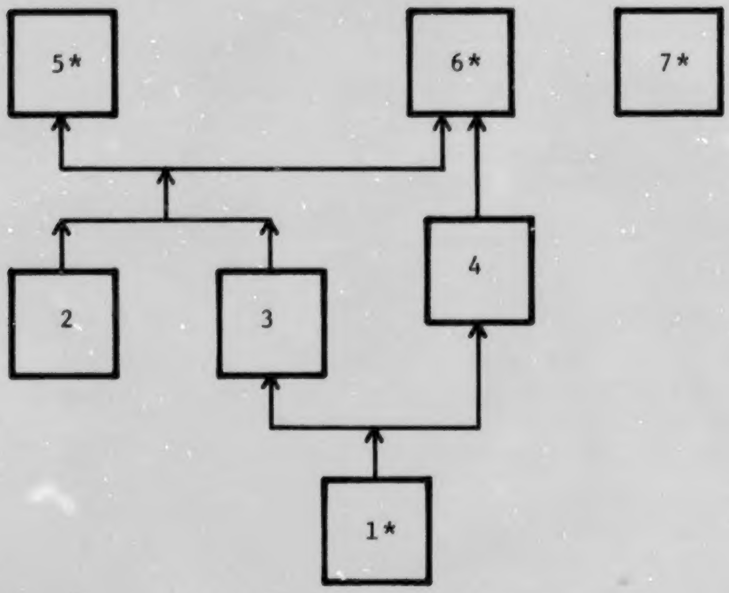
The Adaptive Testing Model

Configuration and Relative Importance of the Objectives

A hierarchical configuration of objectives, such as in Figure 1, defines the interrelationship of the objectives to be mastered by each trainee. Objective 5 has Objectives 2 and 3 as its immediate subordinates or prerequisites. This means that mastery of the skill or competency represented by Objective 5 requires that both Objectives 2 and 3 be mastered. Nonmastery of either or both Objectives 2 and 3 implies nonmastery of Objective 5. The figure indicates no prerequisite to Objective 2. Objective 1 is prerequisite to both Objectives 3 and 4. The immediate prerequisites to Objective 6 are Objectives 2, 3, and 4. No prerequisites are indicated for Objective 7.

Generally, some objectives are considered more important or critical. Other objectives may be subordinate or prerequisite to the former objectives--those of primary concern. If mastery can be ascertained for the "objective of primary concern," then there appears to be little, if any, need to assess performance on the subordinate objectives. If direct assessment of performance on

Figure 1
Hypothetical Hierarchical Configuration of Objectives
(* Indicates an "Objective of Primary Concern.")



all the objectives was desired, then every objective would be identified as an objective of primary concern.

The model assumes that mastery of an objective implies mastery of all its immediate subordinate objectives; nonmastery of an objective implies neither mastery nor nonmastery of the immediate subordinates. Mastery classification on an objective of primary concern results in an assumption that all the immediately prerequisite or subordinate objectives are mastered, unless a subordinate is also of primary concern. Nonmastery classification on an objective of primary concern results in testing each immediate subordinate as if it were also an objective of primary concern.

Basing Decisions on a Data Base

The decisions made in the adaptive testing process are dependent upon information collected from prior examinees. Although the existing model assumes that each prior examinee has answered all the items for each objective, it could accommodate a data base consisting of responses by prior examinees to overlapping subsets of item pools. Decisions such as selection of items for presentation and prediction of correctness/incorrectness of item responses are made on the basis of the interrelation of item responses by prior examinees whose response patterns match the present examinee's pattern. For each item response obtained from an examinee using the adaptive test, a smaller subset of prior subjects' data is used to make decisions--a subset of examinees' dichotomously scored responses exactly like the present examinee's response pattern.

Two response-matching procedures were defined. With the first method a vector $\vec{s}$ of dichotomuously scored responses is generated for an examinee; for each additional response collected within a test, the $\vec{s}$ vector increases. The individual's $\vec{s}$ vector is matched with sets of responses in the data base; but only data base sets with exactly the same $\vec{s}$ vector (the same pattern of "1's" and "0's" to exactly the same questions answered by the examinee) are considered. With the second method, not only is the $\vec{s}$ vector used, but also an $\vec{r}$ vector of mastery/nonmastery classifications for objectives is employed. Only data base sets with exactly the same $\vec{s}$ and $\vec{r}$ vectors are considered. With both methods the matching procedure provides the subset of data base entries that is used for making predictions and selecting other items for presentation.

Predicting item response correctness/incorrectness. Based upon the dichotomously scored responses to items presented to an examinee, conditional probabilities for answering the item correctly or incorrectly are determined on the basis of response patterns in the data base matching the examinee's. If either conditional probability exceeds prespecified levels, the correctness/incorrectness of the examinee's expected response is assumed.

Selection of items for presentation. Based upon an examinee's response pattern and the subset of the data base response matching the examinee's, items that are expected to provide the most information about the objectives of primary concern are selected for presentation. Two selection criteria were investigated in this study: item-objective agreement and inter-item agreement. For each method a coefficient was computed for each item not presented and for which prediction of correctness/incorrectness had not yet occurred. The item with the highest coefficient was presented to the examinee.

For the item-objective method, a coefficient of agreement between item *i* and the *n* objectives of primary concern was calculated as follows:

$$C(i;0_1,0_2,\ldots,0_n|\vec{r},\vec{s}) =$$

$$[\{\sum_{u=1}^{n}[\mathrm{Prob}(0_u = 1)|(\vec{r},\vec{s},i = 1)]\ [\mathrm{Prob}(i = 1)|\vec{r},\vec{s},]\}$$

$$+ [\sum_{u=1}^{n}[\mathrm{Prob}(0_u = 0)|(\vec{r},\vec{s},i = 0)][\mathrm{Prob}(i = 0)|\vec{r},\vec{s}]\}]/n \qquad [1]$$

where

i is the item under consideration;
$0_1, 0_2, \ldots, 0_n$ are the *n* objectives of concern;
$i = 1$ means item *i* is answered correctly;
$i = 0$ means item *i* is answered incorrectly;
$0_u = 1$ means objective *u* is mastered;
$0_u = 0$ means objective *u* is not mastered;

$\vec{r}$ is the vector of objective mastery/nonmastery classifications for the examinee; and
$\vec{s}$ is the vector of the examinee's dichotomously scored item responses.

For the inter-item method a coefficient of agreement between item i and the n other items corresponding to the objectives of concern is computed according to the following formula:

$$A(i;i_1,i_2,\ldots,i_n) = [\{\sum_{j=1}^{n} [\mathrm{Prob}(i_j = 1)|(\vec{r},\vec{s},i = 1)]$$

$$\times [\mathrm{Prob}(i = 1)|\vec{r},\vec{s}]\} + \{\sum_{j=1}^{n} [\mathrm{Prob}(i_j = 0)|(\vec{r},\vec{s},i = 0)]$$

$$\times [\mathrm{Prob}(i = 0)|\vec{r},\vec{s}]\}/n \qquad [2]$$

where

$i_j = 1$ is the probability of answering item i_j correctly;
$i_j = 0$ is the probability of answering item i_j correctly;
$\vec{r}$ is the objective mastery-nonmastery pattern for the examinee; and
$\vec{s}$ is the item response pattern (correct/incorrect) for the examinee.

Examinee response inconsistencies. "Untrue" responses by an examinee are those responses that do not agree with the examinee's "true" response (the examinee's response that is not arrived at by guessing and has not been erroneously selected or created). "Untrue" responses are expected to occur in such cases as

1. Selecting the correct answer by guessing, when in actuality the examinee should have answered the item incorrectly;
2. Providing an incorrect answer because of misinterpretation of part of the question; and
3. Pressing an unintended key on a terminal keyboard.

Item responses that are provided by an examinee, but are contrary to the examinee's "true" response, introduce potential measurement error into any testing process. In the adaptive test model, erroneous responses introduce error into $\vec{s}$, the item response vector. Vector $\vec{s}$ affects predictions of other item responses and selection of items for presentation. Generally, it is expected that item prediction errors will affect the accuracy of the system, whereas errors in item selection will reduce the efficiency of the system. Prediction and selection errors may occur, since the adaptive testing process relies on matching the examinee's $\vec{s}$ with exactly the same response vectors in the data base. Errors introduced into $\vec{s}$ would produce a comparison between the examinee's performance and the wrong subset of prior examinees. Even if some of the response sets in the data base contain the same errors as those made by the present examinee, it would be expected that for each item the majority of prior examinees had provided responses that concur with their "true" responses. Hence, errors introduced into the examinee's item response vector would be expected to compare the examinee's performance to an inappropriate subset of prior examinees.

The adaptive testing model included an optional component that checks for potentially "untrue" responses by comparing the examinee's inter-item response consistency to the inter-item response consistency demonstrated by all prior examinees whose data are included in the data base. When this option was selected, it was necessary that at least two items be presented for the examinee's responses prior to making predictions or to making other item selections based on the item response vector $\vec{s}$. The present model requires that a set of items be independently selected and presented. In this study the number of items presented was sufficient so that the probability of answering all of them correctly by chance alone was less than or equal to .5.

The purpose of obtaining responses to a set of independently selected items was to determine whether the examinee has demonstrated sufficient consistency in his/her response pattern to warrant this pattern serving as the item response vector. A coefficient of relative interrelationship R_x between item x and all other items for which responses have been obtained was computed as follows:

$$R_x = \frac{\sum_i G(x,i)}{\sum_i I(x,i)}, \qquad [3]$$

where

$$G(x,i) = \begin{cases} 1 \text{ if both responses to item } x \text{ and item } i \text{ were correct} \\ \quad \text{or if both responses were incorrect} \\ 0 \text{ if one response was correct and the other was wrong,} \end{cases}$$

and

$$I(x,i) = \{[\Sigma \operatorname{Prob}(i = 1 | x = 1)] \times \operatorname{Prob}(x = 1)\} + \{\operatorname{Prob}(i = 0 | x = 0) \times \operatorname{Prob}(x = 0)\} . \qquad [4]$$

$G(x, i)$ was computed on the basis of the examinee's responses to item x and all the other items presented.

R_x indicates the examinee's consistency as compared to prior examinees' consistency. It is possible that a given examinee demonstrated greater consistency than prior examinees, but when the examinee's consistency was less than that for prior examinees, his/her item response pattern contained "untrue" responses. In this study the criterion for sufficiently consistent responses by an examinee required that for each item x, $R_x \geq .90$. If the criterion was not attained for each item, the item with the lowest R_x value was temporarily removed from consideration as a member of the item response vector $\vec{s}$. Prior to making decisions based on $\vec{s}$, the item response vector must contain at least the required minimum number of elements (equal to the number of items to be answered to insure that the probability of guessing the correct answers is less than the criterion). If $\vec{s}$ contained fewer elements, other items must be independently selected $\vec{s}$. Whenever the number of elements in $\vec{s}$ equaled or exceeded the minimum requirement, item selections and predictions were based upon $\vec{s}$. After the presentation of each additional item, all items for which responses were obtained were included in the calculations of the R_x values. Hence, although an item response may be questioned and not included in $\vec{s}$, a future recalculation may indicate the item response to be consistent with the examinee's other responses. Likewise, items once contained in $\vec{s}$ may be excluded on a future recalculation.

Determining Mastery/Nonmastery of Objectives

For an objective of primary concern, the dichotomously scored results to all its items for which correctness/incorrectness has been determined or predicted were used with the Wald probability ratio test.

For example, suppose that for an objective, responses were obtained to three items and predictions were made for six other item responses. These nine responses (correct/incorrect for each item) were then used in the following formula:

$$S = \left(R \times \log_{10} \frac{C_f}{C_p}\right) + \left[(N - R) \times \left(\log_{10} \frac{1 - C_f}{1 - C_p}\right)\right] \qquad [5]$$

where

R = number of items answered (or predicted as being answered) correctly;
N = number of items (number presented plus the number predicted);
C_f = the critical nonmastery score (difficulty of the objective for nonmasters);
C_p = the critical mastery score (difficulty of the objective for masters).

Mastery/nonmastery classifications were determined by comparing the value of S to ratios involving α and β (Type I and Type II errors); α is the error associated with falsely classifying an examinee as a nonmaster, and β is the error of falsely classifying an examinee as a master:

1. If $S \geq \log_{10} \frac{1-\beta}{\alpha}$, the objective was not mastered.
2. If $S \leq \log_{10} \frac{\alpha}{1-\beta}$, the objective was mastered.
3. If neither of the above conditions was true, no mastery/nonmastery classification was possible (and additional item responses were necessary).

The model assumes that the classification of an objective for which insufficient items exist for a mastery/nonmastery decision is "indeterminate." This decision occurred whenever the pool of available items was exhausted before a mastery/nonmastery decision could be made. Such an objective is presently treated as "unmastered," although this could be altered without affecting other components of the model. Rather than assuming the objective to be unmastered, the process could ascertain which classification zone was approached by the examinee's proportion of items answered correctly. Ferguson (1969) used this procedure, but only after asking for 30 item responses for the objective. It appears that if an examinee cannot demonstrate mastery performance within a realistically expected number of items, immediately prescribing remedial instruction would be more efficient than giving a lengthy test to make a decision. An objective for which an undesirably high proportion of "indeterminate" classifications has been made indicates an insufficient number of items, insufficient item discriminations, or unrealistically high specifications for acceptable misclassification errors.

The adaptive testing procedure terminated when either of the following conditions occurred: (1) all objectives were classified as mastered or unmastered; or (2) the number of prior examinee observations in the data base upon which predictions are based was less than two. For the first condition, the test was terminated. For the second condition, unpresented and unpredicted items corresponding to objectives of concern were randomly presented to the examinee. Termination of the test occurred when each objective was classified.

Eight Versions of the Adaptive Testing Model

The adaptive testing model formulated for this study was applied in a 2 × 2 × 2 configuration of options. These derive from three options, each with two conditions: (1) two methods of item selection based upon item-objective agreement and inter-item agreement; (2) two response matching procedures based upon only item response patterns (only $\vec{s}$) and upon both item response and objective classification patterns (both $\vec{r}$ and $\vec{s}$); and (3) a dichotomous option regarding examinee response inconsistency. Table 1 provides a delineation of the options used for each version; the numbers used in the remainder of the report refer to combinations of options employed.

Phase I: Monte Carlo Simulations

The purpose of this phase of the study was twofold: (1) to test for the relative accuracy and efficiency of the eight versions of the adaptive testing model and a control version and (2) to study the relation of loss to individuals' achievement levels for the adaptive testing versions. Accuracy was examined in terms of correct mastery/nonmastery classifications. Efficiency was investigated in terms of the number of items presented to examinees.

The control version to which the adaptive testing versions were compared involved the testing of every objective. For each objective a prespecified number of items was randomly selected for each examinee. Under the control treatment, examinees generally received different items for an objective, but each received the same number of items. For each objective a randomly selected integer between 3 and 6, inclusive, was chosen for the number of items to be presented. Mastery of an objective was obtained if an examinee obtained a score of N-1 or higher, where N equals the number of items presented. A score of less

Table 1
Options Employed in the Eight Versions
of the Adaptive Testing Model

Testing Version	Item Selection Method	Response Matching[1] Procedure	Inconsistency Check
1	Item-objective	Only $\underline{\vec{s}}$	No
2	Inter-item	Only $\underline{\vec{s}}$	No
3	Item-objective	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	No
4	Inter-item	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	No
5	Item-objective	Only $\underline{\vec{s}}$	Yes
6	Inter-item	Only $\underline{\vec{s}}$	Yes
7	Item-objective	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	Yes
8	Inter-item	Both $\underline{\vec{r}}$ and $\underline{\vec{s}}$	Yes

[1] $\underline{\vec{s}}$ is the item response vector and $\underline{\vec{r}}$ is the objective mastery/nonmastery classification vector.

than N-1 resulted in a nonmastery classification. The resulting lengths of the tests and the mastery criteria reflected the parameters used in the Air Force Weapons Mechanics training program at Lowry Air Force Base, Denver, Colorado.

Item response generation. Item response data were generated for hypothetical examinees who were to demonstrate some consistency in performance across examinations. This assumes that individuals in instructional programs demonstrate a certain consistent performance in mastering or not mastering objectives.

For each examination by adaptive test version, two sets of examinee data were generated--one representing past examinees' responses and the other including responses that would be obtained from present examinees. For the control version, only one set of data was generated for each examination. A set of examinee responses was generated in two steps using two computer programs, GENTAB and GENRESP. For each examinee GENTAB produced values for elements of consistency to be demonstrated across testings. These elements were the examinee's achievement level and risk of guessing. The values from GENTAB and additional parameters were used to produce item responses through program GENRESP. Parameters specified for GENRESP included the following: (1) hierarchical configuration of the objectives; (2) objective parameters, such as difficulty; (3) discrimination, and passing criteria; (4) proportion and type of hierarchical errors; and (5) guessing factor for answering items correctly.

Generation of examinees' true item responses. For each objective, each item response for an examinee was based on a probability of answering the item correctly. The algorithm used was

$$P\ (u = 1) = \begin{cases} d + \dfrac{\theta - \overline{\theta}}{1 - \overline{\theta}}\,(1 - d) \text{ if } \theta \geq \overline{\theta} \\[2ex] d + \dfrac{\theta - \overline{\theta}}{\overline{\theta}}\, d \text{ if } \theta < \overline{\theta}\ , \end{cases} \qquad [6]$$

where

$P(u = 1)$ = the probability of answering the item correctly;
d = difficulty of the item;
$\overline{\theta}$ = examinee's objective score; and
θ = mean objective score of the corresponding mastery/nonmastery group.

A random number r in the closed interval 0 to 1 was selected. If $r \leq P(u = 1)$, the examinee was assigned a correct item response; otherwise, an incorrect item response was assigned.

Inclusion of examinee error. The factor of successful guessing was included in GENRESP. The probability that an examinee would attempt to guess the correct answer, given that his/her "true" response would be incorrect, was derived by the formula

$$P_1 = g_1(1 - \theta d) \qquad [7]$$

where

g_1 is the risk factor for the examinee (from GENTAB);
θ is the examinee's objective score; and
d is the item difficulty for the examinee's mastery or nonmastery group.

A random number r in the interval 0 to 1 was selected. If $r_1 \leq P_1$, the examinee would attempt to guess the correct answer. The probability of guessing correctly was obtained from the formula

$$P_2 = g_2 + g_2\ \theta d \qquad [8]$$

where g_2 is the guessing factor for the item (the probability of randomly selecting the correct answer), and θ and d are the same as defined previously. For all items, g_2 was set equal to .2, assuming five alternatives to each item. A random number r_2 in the interval 0 to 1 was selected. If $r_2 \leq P_2$, the examinee was credited with answering the item correctly.

Experimental Design

The design employed 90 cells comprised of an element from each of the following two dimensions (independent variables): (1) Testing Version (8 adaptive test versions and 1 control test version) and (2) Examination (10 examinations). For each testing version, data were simulated for 50 hypothetical examinees, each of whom was to take 10 examinations using only 1 testing version across the 10 examinations. Hence, there were 450 hypothetical examinees, each taking 10 examinations.

Separate split-plot factorial analyses of variance were conducted for each of two dependent variables. The dependent variables were (1) total loss associated with errors in mastery/nonmastery classifications and (2) total number of items presented.

Total loss. A loss value is a positive or zero number assigned to an action-outcome combination (Hays & Winkler, 1970). A zero loss value is assigned to any combination that reflects the best actions under the true circumstances. If an action is less desirable than the best actions, an error is associated with the action and is assigned a positive value reflecting the level of error involved.

The loss values appearing in Table 2 represent the relative amounts of loss attributed to each mastery/nonmastery/indeterminate decision made, given the "true" mastery/nonmastery status.[1] It can be seen in Table 2 that under the known true situation of mastery, the best decision was to classify performance on an objective as "mastery." The positive numbers for decisions of "nonmastery" and "indeterminable" indicate there were errors involved with these decisions--the greater error being associated with the latter. Total loss equals the sum of the separate losses incurred for each objective decision for an examinee.

Table 2
Matrix of Loss Values Provided for
Objectives of Primary and Secondary Concern

	True Classification	
Classification Decision	Mastery	Nonmastery
Objectives of Primary Concern		
Mastery	0	10
Nonmastery	5	0
Indeterminable	7	3
Objectives of Secondary Concern		
Mastery	0	6
Nonmastery	4	0
Indeterminable	5	2

Total number of items presented. Items for the adaptive tests were presented to provide information for predicting correctness/incorrectness of other items. The total number of items presented refers to the number of items answered by an examinee in order to make mastery/nonmastery decisions on objectives.

---

[1] Roger Pennell of the Air Force Human Resources Laboratory at Lowry Air Force Base provided losses based upon values independently obtained from individuals knowledgeable of the Air Force Weapons Mechanics training program.

Experimental model. The split-plot factorial model used was

$$X_{ijkm} = \mu + A_i + B_j + \pi_{k(i)} + AB_{ij} + B\pi_{jk(i)} + \varepsilon_{m(ijk)} \qquad [9]$$

where

X_{ijkm} is the dependent variable;
A_i is the testing version;
B_j is the examination; and
$\pi_{k(i)}$ is the subject effect.

A posteriori tests. With regard to the testing version effect, the Dunnett's *t* statistic was computed for each adaptive testing version with the control treatment. This a posteriori test was used for each dependent variable, regardless of the F value obtained using the analysis of variance (Winer, 1971, p. 201). Therefore, each version was compared with the control treatment. For other effects, Newman-Keuls tests were performed only when significant F values ($\alpha = .05$) were obtained from the analyses of variance.

Sample size. Each data base from which predictions were made was composed of 300 sets of responses. For each of the 90 testing versions by examination cells, 50 hypothetical examinees were used.

α and β levels. In this phase of the study, the values of α and β relative to the Wald procedure were set at .2 and .1, respectively.

Results

All of the adaptive testing versions were significantly more efficient than the control version. Only one adaptive testing version demonstrated significantly smaller losses than the control version. An analysis of variance indicated significant Examination and Testing Version × Examination effects ($\alpha = .05$). A quasi-F statistic was computed for the testing version, since the mixed-effects model did not directly provide a mean sums-of-squares estimate for the required denominator (Winer, 1971, pp. 375-378). Table 3 shows the results of the analysis of variance, and Table 4 provides the descriptive statistics for each testing version.

The use of Hartley's test for homogeneity of variance (Winer, 1971, pp. 207-208) resulted in a rejection of the equal variance assumption. Hence, a more conservative test proposed by Box (Winer, 1971, p. 206) was used. The degrees of freedom corresponding to each numerator were reduced to one. The test effect remained significant at the .05 level, but the Treatment × Test interaction did not.

Dunnett's test indicated that the only adaptive testing version signifi-

Table 3
Analysis of Variance For Total Loss

Source	*df*	Mean Square	F
Between Subjects	449		
Testing version	8	623.32	1.14
Subjects-within-groups	441	525.15	
Estimates for quasi-F calculations	457	544.624	
Within Subjects	4050		
Examination	9	754.74	36.61*
Testing version × examination	72	40.09	1.94**
Examination × subjects-within-groups	3969	20.62	

*p <.01.
**p <.01 for *df*(72,3969); p <.25 for *df*(1,3969).

cantly different (α = .05) from the control test was the sixth version--Adaptive Testing Version 6, using the inter-item agreement, based only on the item response vector, and employing the inconsistency check. Although the obtained *t* value of Adaptive Testing Version 7 did not exceed the critical value, the difference in the two was extremely small. The losses obtained for both versions were extremely close. Adaptive Testing Version 7 used item-objective agreement, based on both item response and objective classification vectors, and employed the inconsistency check.

Table 4
Descriptive Statistics of Total Loss for
Each Examinee per Testing Version

Testing Version	Mean	SD	Range Min	Range Max
1	5.84	10.25	0	52
2	5.52	9.56	0	48
3	5.88	9.79	0	52
4	5.39	9.25	0	60
5	5.03	8.46	0	46
6	4.73	8.63	0	49
7	4.84	8.55	0	50
8	5.05	8.40	0	44
Control	8.40	8.45	0	60

The Newman-Keuls test indicated no pattern of significantly different losses among the examinations. Although significant differences did occur between some pairs of examinations, no trend was indicated. The Testing Version × Examination interaction was not significant using the conservative F test. There was a tendency for all versions of the model to obtain approximately the same losses for each examination and to have losses less than the conventional test, except for the third examination.

For the number of items presented, an analysis of variance indicated significant Testing Version, Examination, and Testing Version × Examination effects ($\alpha = .05$). As with the other dependent variables, a quasi-F statistic was calculated for the testing version effect. All the effects were also significant ($\alpha = .05$) for the more conservative F test, used because of the heterogeneous variances. Table 5 shows the results of the analysis, and Table 6 provides the descriptive statistics for the number of items presented.

Table 5
Analysis of Variance For Number of Items Presented

Source	df	Mean Square	F
Between Subjects	449		
Testing version	8	31285.58	256.06*
Subjects-within-groups	441	2.56	
Estimates for quasi-F calculations	72	122.18	
Within Subjects	4050		
Examination	9	276.51	142.53*
Testing version × examination	72	121.56	62.66*
Examination × subjects-within-groups	3969	1.94	

*$p < .01$.

The results of the Newman-Keuls tests for the testing version effect showed that each adaptive test required significantly fewer ($\alpha = .05$) items than the control test. There were no significant differences among the adaptive versions.

Although significant differences existed in numbers of items presented for the 10 examinations, the adaptive testing versions varied only slightly in their relative efficiency. A version that appeared to require the fewest items on one examination may have required the most on another examination. The differences in the number of items required by the adaptive versions for any one test were not substantially different.

Loss as a Function of Achievement Levels

Although Adaptive Testing Version 6 demonstrated overall superior accuracy, the losses incurred for all examinees were not the same. More importantly, the losses relative to examinees' general achievement may be small for some levels but high for others. The mean losses as a function of examinees' achievement

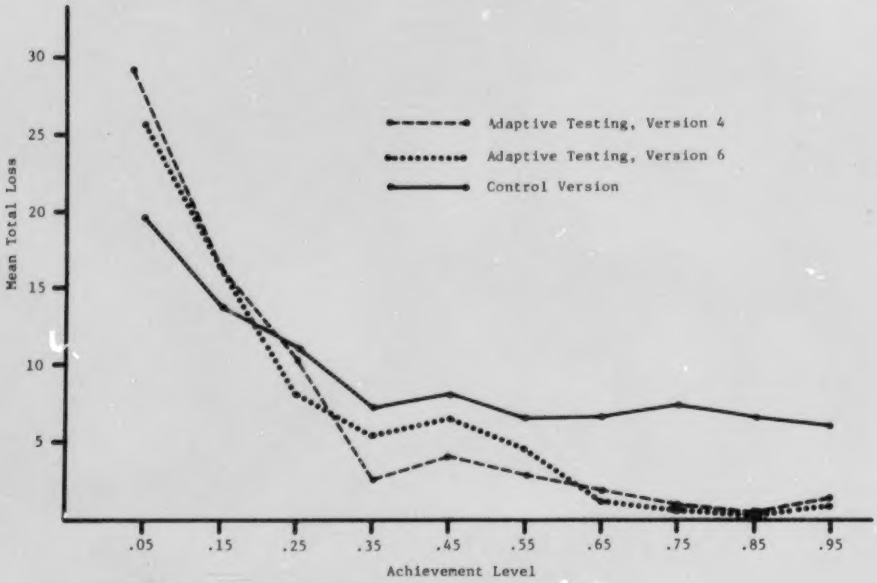
Table 6
Descriptive Statistics for Number of Items Presented

Testing Version	Mean	SD	Range Min	Range Max
1	2.88	1.50	2	11
2	3.09	1.72	2	9
3	2.92	1.38	2	7
4	2.84	1.30	2	8
5	3.45	1.60	2	12
6	3.48	1.92	2	14
7	3.47	1.90	2	15
8	3.34	1.64	2	14
Control	26.90	4.35	20	34

levels are shown for Adaptive Testing Versions 4 and 6 and for the control testing version in Figure 2.

The comparison of losses with respect to achievement levels demonstrated that both adaptive testing versions performed equally well throughout the achievement range. Version 6 demonstrated a slight advantage over Version 4 in the lower end of the achievement levels. This was probably due to the consistency check employed in Version 6.

Figure 2
Mean Total Loss Corresponding to Levels of Achievement



The adaptive testing versions had smaller losses for the middle and upper achievement levels, but this was reversed for the lower levels. This difference could be eliminated by reducing the α level. It may be recalled that β was set to .1, whereas α was set at .2. Since the false nonmastery error would be larger than the false mastery error, a higher proportion of false classifications would be expected for those at the lower achievement levels.

The adaptive testing versions may have produced more inaccurate classifications due to the paucity of data representative of poorer-achieving students. Since only a small proportion of examinees in the data base did not master the objectives, the predictions made for the poorer-achieving students were often based on relatively few data cases. Such was not the case for those with higher achievement levels.

Selection of Adaptive Testing Versions for the Next Phase

The intention of the next phase of the study was to compare the results of some of the adaptive testing versions with those obtained in the present testing system used in the Air Force Weapons Mechanics training program at Lowry Air Force Base. Adaptive Testing Versions 4 and 6 were selected. No version was significantly superior in numbers of items presented. Adaptive Testing Version 6 was selected because of its superior accuracy. Adaptive Testing Version 4 was selected, however, solely on the basis of the mean number of items presented for item prediction.

Phase II: Real Data Simulations

Purpose

The purpose of this phase of the study was to compare (1) the relative efficiency of Adaptive Testing Versions 4 and 6 with each other and with the present testing method used in the Weapons Mechanics training program and (2) the classification decisions made from the adaptive testing versions with those made by the pesent method used in the Weapon Mechanics training program.

Design

The control testing version for this phase was a testing procedure consisting of a fixed set of items for each objective. Hence, all examinees answered the same set of items under the control treatment.

Classification decisions made by the adaptive testing and control testing versions were compared using an index defined as the number of agreements minus the number of disagreements. An agreement in classifying an examinee's performance on an objective was obtained when both indicate "nonmastery." Since for the adaptive tests, performance classified as "indeterminate" dictated procedures identical to those classified as "nonmastery," this condition was also considered an agreement. The α and β values selected were the same as in the previous phase--.2 and .1, respectively.

Data that were actually collected on four examinations in the Weapon Mechanics training program were used in the computer simulations for this phase.

For each examination, from 250 to 290 response sets were available. It was not feasible to match student identification codes across the examinations, since there was no control over the forms of the tests taken by the examinees. For each examination, the first 150 response sets, sorted in ascending chronological order, were used to form the data base. Of the remaining subjects, 50 were randomly selected as the examinees who were to take the simulated adaptive tests. Hence, within each examination the same 50 trainees were used as examinees, regardless of the testing version; but the same 50 trainees were not used across examinations.

The assumed hierarchical configurations for the objectives for each examination were provided by Roger Pennell of the Air Force Human Resources Laboratory, Lowry Air Force Base, Denver, Colorado. The mastery score for an objective with $N(\geq 2)$ items was set to $N - 1$, as is presently done with conventional testing procedure, which is referred to here as the control testing version. If N equaled 1, the cutting score was set to 1.

Correlated t tests were used to compare adaptive testing versions. A t test for a mean equal to a constant was employed for each comparison of each adaptive testing version to the control testing version.

Results

Both adaptive testing versions used in this phase of the study demonstrated that each required significantly fewer items than the control testing version. Version 4 of the model required the presentation of fewer items than Version 6.

Efficiency. Adaptive Testing Version 4 required statistically significantly ($t = 8.30$, $df = 199$, $p < .001$) fewer items than Version 6. The descriptive statistics for these versions are shown in Table 7. Although there was a statistical difference, the superior efficiency of Version 4 amounted to less than one item per examinee per examination.

Table 7
Descriptive Statistics for Adaptive
Testing Versions 4 and 6

Variable and Statistic	Adaptive Testing Version 4	Adaptive Testing Version 6
Number of Items Presented		
Mean	3.02	3.92
SD	1.19	1.42
Index of Agreement		
Mean	6.15	5.54
SD	3.39	3.27

Mastery/nonmastery decisions. Adaptive Testing Version 4 had a statistically significantly ($t = 5.58$, $df = 199$, $p < .001$) higher agreement in mastery/nonmastery classifications than Version 6. The descriptive statistics for these versions are also shown in Table 7.

The average number of objectives per examination was 7.25. Hence, the range of the index could be from -7.25 to 7.25. A complete agreement in decisions would result in an index value of 7.25; a complete disagreement would result in a value of -7.25. In terms of percent of agreements in decisions, Versions 4 and 6 had 92% and 88% agreement with the control testing version, respectively.

Separate _t_ tests were performed on the number of items presented for each of the adaptive testing versions compared to the number required by that of the control version. The mean number of items presented under the control testing version across the four tests was 15.25. The number of items required by the adaptive testing version are presented in Table 8. The visual comparison of the tabled values reveals such large differences that no statistical test was necessary.

Since the four examinations differed in hierarchical configurations, number of objectives, and number of available items, Table 8 presents the percent of reduction in test items required by the adaptive testing versions in relation to the control testing version for each examination. The table also shows the percent of agreements in mastery/nonmastery decisions between each adaptive testing version and the control testing version.

Table 8
Comparison of Results of Adaptive Testing Versions 4 and 6 to Control Testing Version for Each Examination

Adaptive Testing Version and Examination	Number of Items Presented: Control Version	Number of Items Presented: Adaptive Testing Version	Percent of Item Reduction	Number of Objectives	Percent of Mastery and Non-Mastery Agreements
Version 4					
1	20	4.3	79	14	91
2	12	2.6	78	4	98
3	14	2.5	82	6	86
4	15	3.2	79	5	99
Version 6					
1	20	5.1	75	14	87
2	12	4.2	65	4	92
3	14	2.4	83	6	84
4	15	4.0	73	5	93

The results show that both Adaptive Testing Versions 4 and 6 made most of the same mastery/nonmastery decisions as were presently being made by the Air Force in its Weapons Mechanics program; but the adaptive testing versions make the decisions with approximately 75% fewer items than the conventional, or control, version.

Discussion and Conclusions

Both simulation phases of the study have shown that the adaptive testing versions could make mastery/nonmastery decisions much more efficiently than testing on each objective with a constant number of items for each objective presented.

The real-data simulation showed that the mastery/nonmastery agreement between the control testing version and the adaptive testing versions was higher for Adaptive Testing Version 4. This does not mean that Version 4 is more accurate than Version 6. On the contrary, in the first simulation it was demonstrated that Adaptive Testing Version 6 was the only adaptive procedure that had significantly smaller loss than the control version. In essence, Adaptive Testing Version 4 and the control version in the second simulation phase would be expected to be equally as inaccurate in mastery/nonmastery decisions. Adaptive Testing Version 6 would be expected to be more accurate than the control version and, hence, would have fewer agreements with the control version than would Version 4.

Although in both phases statistically significant differences were found among the adaptive testing versions, the assignment of different values to the version's parameters might equalize all results. All the adaptive testing versions were used with the same values specified for the model's parameters. For example, for all versions, α and β were set at .2 and .1, respectively. The versions may be differentially sensitive to the parameters. Hence, two versions may be expected to perform exactly the same, but only by specifying different values for the same parameters.

For both simulation phases of the study the number of sets of responses needed in the data bases were unknown. For the second simulation phase it was estimated that 150 sets would be sufficient. The results indicate that an average of 29 sets matched each examinee's set on each test. The average of 29 sets per examinee did not give sufficient information as to whether the data base was of sufficient size. The ranges in number of sets indicated that for every test and for every adaptive testing procedure the data base was completely depleted for some examinees. As in the first simulation, it may not be that the data base contained insufficient numbers of response patterns but that there was an insufficient number of patterns for poorer performing individuals. In both phases the data bases were composed of response patterns representative in type and proportion to those patterns expected in the population of examinees. It appears that when a high proportion of examinees mastered the objectives, as in the Weapons Mechanics program, such a data base is insufficient for predictions of performance by nonmastering examinees. Hence, in such a situation, oversampling of nonmastering examinees may be required in order to provide adequate data for all levels of performance.

Because of the similarity of the results for all the adaptive testing versions in the monte carlo simulations and the superior efficiency demonstrated by Adaptive Testing Versions 4 and 6 procedures in the real-data simulations, it appears that any of the adaptive testing variations used in this study would be much more efficient than the conventional testing procedure used by the Air Force.

REFERENCES

Ferguson, R. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Ferguson, R. A model for computer-assisted criterion-referenced measurement. Education, 1970, 91, 25-31.

Hays, W., & Winkler, R. Statistics, Volume 1: Probability, inference, and decision. New York: Holt, Rinehart, & Winston, 1970.

Kalisch, S. J. A tailored testing model employing the beta distribution and conditional difficulties. Journal of Computer-Based Instruction, 1974, 1, 22-28. (a)

Kalisch, S. J. The comparison of two tailored testing models and the effects of the models' variables on actual loss. Unpublished doctoral dissertation, Florida State University, 1974. (b)

Wald, A. Sequential analysis. New York: Dover Publications, 1973. (Originally published, 1947.)

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

ACKNOWLEDGMENTS

This study was sponsored by the Air Force Human Resources Laboratory, Air Force Systems Command, United States Air Force, Brooks AFB, San Antonio, TX 78235.

A COMPARISON OF ICC-BASED ADAPTIVE MASTERY TESTING AND THE WALDIAN PROBABILITY RATIO METHOD

G. GAGE KINGSBURY AND DAVID J. WEISS
UNIVERSITY OF MINNESOTA

The use of criterion-referenced achievement test interpretation has gained great support within the educational measurement community since its introduction less than two decades ago (Glaser & Klaus, 1962). It is intuitively appealing to educators to be able to measure students' performances against an absolute standard of behavior on prespecified learning objectives, and the use of criterion-referenced test interpretation gives educators this capability. One of the most basic forms of criterion-referenced test interpretation involves classifying students into two categories--one containing students who have achieved a sufficient command of the subject matter (mastery) and the other containing students who have not achieved a sufficient command of the subject matter (nonmastery). Traditionally, a student is declared a master if his/her score on a conventional classroom achievement test is as high or higher than a prespecified cutoff point or is declared a nonmaster if his/her score on the test is lower than the cutoff point. This form of classroom testing has been called mastery testing and can be useful (1) in determining the degree of student proficiency within a classroom and (2) as a diagnostic tool to identify individuals who need further training in specific instructional areas (Nitko & Hsu, 1974).

As traditional mastery testing has been developing its own technology, adaptive testing technology has also developed to allow educators to make maximum use of classroom testing time while reducing the amount of time spent on testing to a minimum. The use of adaptive testing techniques has recently been shown to be effective in reducing test length while obtaining high-fidelity achievement level estimates in several instructional settings (e.g., Bejar, Weiss, & Gialluca, 1977; Brown & Weiss, 1977).

Mastery and adaptive testing technologies have each shown their usefulness in the academic setting for different, but compatible, reasons. It is therefore not surprising that a fusion of the two techniques should occur in order to allow mastery testing to be accomplished in the shortest possible class time while maintaining the accurate decisions necessary for correct diagnoses of student instructional problems.

Approaches to Adaptive Mastery Testing

Two attempts that have been made to combine mastery and adaptive testing technologies have been Ferguson's (1969, 1970) application of Wald's Sequential

Probability Ratio Test (SPRT) to mastery testing and Kingsbury and Weiss's (1979a) formulation of an item characteristic curve (ICC) approach to adaptive mastery testing (AMT). Both of these testing procedures attempt to accomplish two common ends. First, the procedures seek to shorten the length of the test. Second, the procedures use statistical techniques designed to hold the number of misclassifications (i.e., individuals for whom the wrong decision is made) to some acceptable minimum. The methods by which these two procedures attempt to accomplish these ends are quite different.

The very fact that two procedures exist that attempt to accomplish the same basic ends through different techniques renders a comparison of the two methods desirable. The prime objective of this paper, then, was a comparison of the efficiency with which these two procedures for mastery testing achieved their goals of reducing test length while obtaining a high percentage of correct decisions. The first level of comparison presented here is a descriptive comparison based on the theories underlying each of the procedures. This is followed by an empirical comparison of the two testing procedures within the context of a monte carlo simulation of test responses designed to fit a number of theoretical contingencies.

Wald's SPRT Applied to Mastery Testing

The SPRT procedure. Wald's (1947) SPRT was originally designed as a quality control test for use in a manufacturing setting. It was designed to determine whether a large consignment of products (e.g., light bulbs) contained a small enough proportion of defective bulbs to pass some prespecified quality criterion while only testing a small sample of the light bulbs in the consignment. Wald's solution to this problem was to draw light bulbs sequentially from the consignment, to test the light bulb drawn at each stage, and to determine at each stage the relative probabilities of the following two hypotheses:

$$H_0: p = p_0 \qquad [1]$$

$$H_1: p = p_1 \qquad [2]$$

where

p = the proportion of defective elements (light bulbs) in the population (consignment);

p_0 = the proportion of defective elements in the population below which it is always desired to accept the quality of the population; and

p_1 = the proportion of defective elements in the population above which it is always desired to reject the quality of the population.

Since each stage of the sampling procedure may be viewed as a Bernoulli trial (given that each element is sampled at random without replacement from the population of equivalent elements and assigned either nondefective or defective status), the probability of observing a certain number of defective elements in a sample of a certain size, given that either H_0 or H_1 is true, may be described with the binomial probability function. Consequently, the probability of observing W defective elements in a sample of m elements (W_m), under H_0: $p = p_0$ is

$$p_{0_m} = p^{(m-W_m)} (1 - p_0)^{W_m} . \qquad [3]$$

Under H_1: $\underline{p} = \underline{p}_1$, the probability becomes

$$p_{1_m} = p_1^{(m-W_m)} (1 - p_1)^{W_m} . \qquad [4]$$

The ratio of these two probabilities yields an index of the relative strengths of the two hypotheses such that at each stage in the sampling procedure the quality of the consignment may be either rejected or accepted, or sampling of elements may be continued. The stringency of the test is based (1) on the proportion (α) of errors willing to be tolerated in rejecting the quality of the consignments that actually do have the quality desired and (2) on the proportion (β) of errors willing to be tolerated in accepting the quality of consignments that do not actually have the minimum acceptable quality.

In its final log form the test used by the SPRT at each stage of sampling specifices that if

$$\text{Log} \frac{p_{1_m}}{p_{0_m}} \geq \text{Log} \frac{1 - \beta}{\alpha} , \qquad [5]$$

the consignment is rejected; if

$$\text{Log} \frac{p_{1_m}}{p_{0_m}} \leq \text{Log} \frac{\beta}{1 - \alpha} , \qquad [6]$$

the consignment is accepted; and if

$$\text{Log} \frac{\beta}{1 - \alpha} < \text{Log} \frac{p_{1_m}}{p_{0_m}} < \text{Log} \frac{1 - \beta}{\alpha} , \qquad [7]$$

sampling continues.

Wald (1947) has shown that this testing procedure results in error levels approximating α and β across consignments. Further, it has been shown that the probability of not obtaining a decision for a consignment approaches zero as the sample size increases.

Ferguson's application to mastery testing. Ferguson (1969) has applied the SPRT within a mastery testing situation using test item responses in place of light bulbs and a domain of items that represents an instructional objective instead of a consignment. The quality that Ferguson evaluated was students' command of the content area being tested. Ferguson also branched through an instructional hierarchy, applying the SPRT to various objectives of instruction. The present study, however, will concentrate on the application of SPRT to a single instructional unit.

To employ the SPRT in a mastery testing situation, the educator must specify the following:

1. Two criteria of performance ($\underline{p}_0$ and $\underline{p}_1$), which serve as the lowest level at which a mastery decision will be made and the highest level at which a nonmastery decision will be made and which bound the uncertainty region in which testing will continue.

2. Two levels of error acceptance (α and β), which determine the strictness of the decision test and should reflect the relative costs of the two error types.

3. A maximum test length to constrain the testing time for individuals who are very difficult to classify.

One characteristic of this form of adaptive mastery testing is that it is fairly simple to implement within a classroom situation. The decision rule is easily incorporated into a chart that shows the teacher or the student how many questions need to be answered correctly or incorrectly for each test length in order to terminate the test. Once the charts are made for various values of $\underline{p}_0$, $\underline{p}_1$, α, and β, the statistical work is completed. This puts the power of the SPRT procedure into the hands of the educator quite readily. The procedure is not fully adaptive, however. Items are selected at random or in a fixed sequence; it is only the test length that varies for individuals.

ICC-Based Adaptive Mastery Testing (AMT)

The paradigm for AMT that Kingsbury and Weiss (1979) have proposed makes use of ICC theory and Bayesian statistical theory to adapt the mastery test to the individual´s level of skill during the testing process. ICC theory is used to estimate the parameters that most efficiently describe each of the items in the item pool. Given these parameter estimates, it is possible to prescribe a type of adaptive procedure that may allow mastery decisions that are quite accurate to be made while shortening the length of the test needed for most individuals.

The AMT procedure is based on three integrated procedures. These are (1) a procedure for individualizing the administration of test items, (2) a method for converting a traditional (proportion correct) mastery level to the latent achievement metric, and (3) a procedure for making mastery decisions using Bayesian confidence intervals.

Individualized item selection. To make mastery testing a more efficient process, it is desirable to reduce the length of each individual´s test (1) by eliminating test items that provide little information concerning an individual´s achievement level and (2) by terminating the AMT procedure after enough information has been gathered so that the mastery decision can be made with a high degree of confidence. To operationalize this goal, an item to be administered to an individual at any point during the testing procedure is selected on the basis of the amount of information that the item provides concerning the individual´s achievement level estimate at that point in the test, since that item should provide the most efficient use of testing time. A procedure that selects and administers the most informative item at each point in an adaptive test--the maximum information search and selection (MISS) technique--has been described by Brown and Weiss (1977) and is part of the AMT procedure.

The information that an item provides at each point along the achievement continuum may be determined using the ICC model that is assumed to underly individuals' responses to test items. The AMT procedure assumes the 3-parameter logistic ICC model (Birnbaum, 1968). Using this model, the information available in any item is (Birnbaum, 1968, Equation 20.4.16)

$$I_i(\theta) = (1 - c_i)D^2 a_i^2 \psi^2 \,[DL_i(\theta)] \;/\; \{\psi[DL_i(\theta)] + c_i \psi^2\,[-DL_i(\theta)] \;, \qquad [8]$$

where

$I_i(\theta)$ = the information available from item i at any achievement level, θ;
c_i = the lower asymptote of the ICC for the item;
D = 1.7, a scaling factor used to allow the logistic ICC to closely approximate a normal ogive;
a_i = the discriminatory power of the item at the inflection point of the ICC;
Ψ = the logistic probability density function;
$L_i(\theta)$ = $a_i(\theta - b_i)$ where b_i is the difficulty of the item; and
Ψ = the cumulative logistic function.

If it is assumed that the achievement level estimate ($\hat{\theta}$) is the best estimate of the actual achievement level (θ), the item information of each of the items not yet administered may be evaluated at $\hat{\theta}$ at any point during the test. The item that has the highest information value at the individual's current level of $\hat{\theta}$ is thus chosen to be administered next.

For this study a Bayesian estimator of the individual's achievement level, developed by Owen (1969), was used. This estimation procedure has been shown to yield biased estimates of trait levels (Kingsbury & Weiss, 1979; McBride & Weiss, 1976). This bias may be attributed to the assumption of a normal distribution of θ in the population made by Owen's procedure or due to inappropriate prior information concerning θ on the individual level (Kingsbury & Weiss, 1979b). The bias inherent in this scoring strategy may render the MISS technique less efficient than it would be under optimal conditions, thereby reducing the efficiency of the AMT technique as a whole.

To use MISS under optimal conditions, trait level estimates should be obtained by maximum likelihood estimation, which yields asymptotically efficient estimates (Birnbaum, 1968). Maximum likelihood estimation techniques are not able, however, to obtain trait level estimates for consistent item response patterns (either all correct or all incorrect) or for item response patterns for which the likelihood function is extremely flat. The Bayesian technique will yield an estimate for any response vector. This inability to estimate θ for some response patterns mitigated against the use of a maximum likelihood estimation procedure for AMT. Consequently, the Bayesian estimation procedure was used in the AMT procedure on the assumption that the capability to obtain a θ estimate for each individual at each point during the test would outweigh any efficiency lost due to the bias inherent in the estimation procedure. The use of the Bayesian estimation strategy in this study also allowed the use of easily interpretable Bayesian confidence intervals to make the mastery decision.

Mastery level. The classical mastery testing procedure specifies a percentage of the items on a test that must be correctly answered by an individual in order for him/her to be declared a master. Using ICC theory, it is possible to generate an analog to the percentage cutoff of classical theory for use in adaptive testing, even though the use of MISS will tend to result in each person answering about 50% of the items correctly, given a large enough item pool (because items administered will most probably be close to the individual's level of θ). The analog is based on the use of the test characteristic curve (TCC; Lord & Novick, 1968). The TCC is the function that relates the achievement continuum to the expected proportion of correct answers that a person at any level of θ may be expected to obtain if all of the items on the test are administered.

For this procedure the assumption was made that a 3-parameter logistic ogive described the functional relationship between the latent trait (achievement) and the probability of observing a correct response to any of the items on the test. This assumption yields a TCC of the following form:

$$E(P|\theta) = \sum_{i=1}^{n} (1 - c_i) + c_i \frac{1 + \exp[1.7a_i(b_i - \theta)]}{\exp[1.7a_i(b_i - \theta)]} \Bigg/ n \qquad [9]$$

where

$E(P|\theta)$ = the expected value of the proportion of correct answers observed on the test given at any achievement level;

$\underline{n}$ = the number of items on the test;

$\underline{c_i}$ = the estimate of the lower asymptote for the ICC of item $\underline{i}$;

$\underline{a_i}$ = the estimate of the discriminatory power for the item;

$\underline{b_i}$ = the estimate of the difficulty of the item; and

θ = any given achievement level.

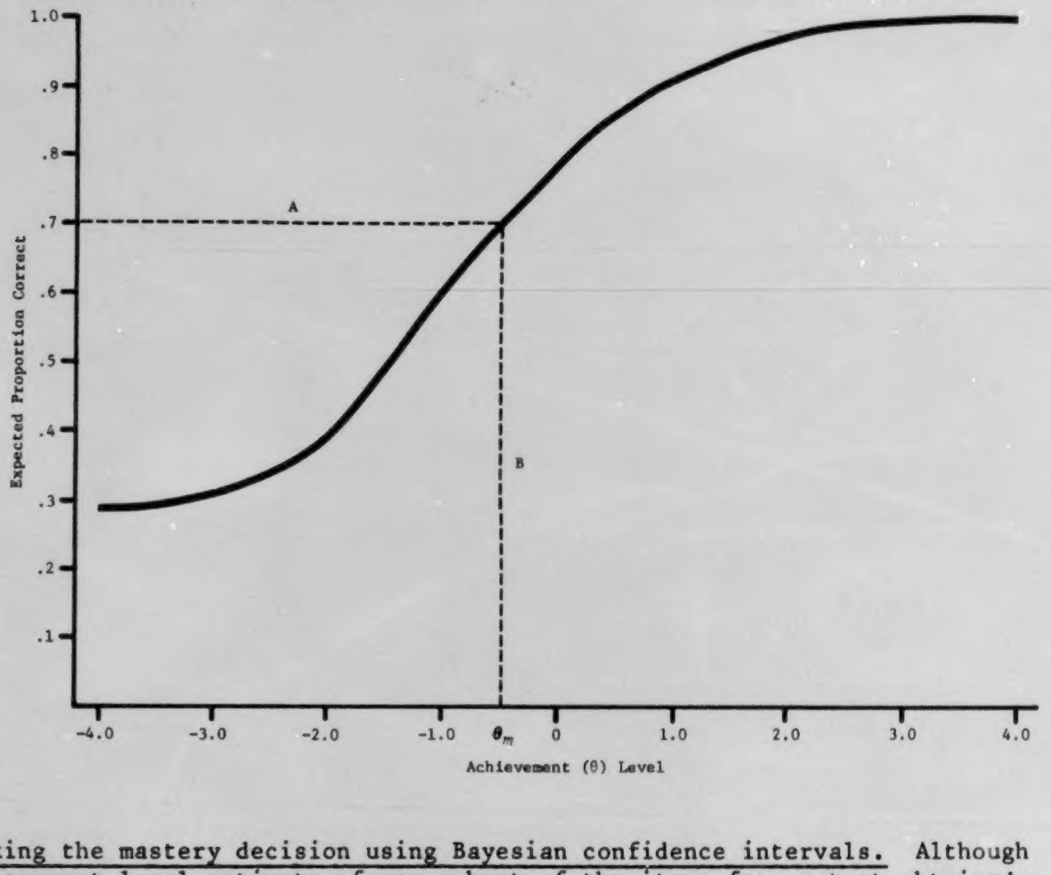
This monotonically increasing function enables the expression of any given level of θ to its most likely proportion correct or, more importantly in this context, to determine the level of θ that will most probably result in any given proportion of correct answers. To exemplify the use of the TCC in determining a level of θ that is comparable to a desired percentage mastery level, a hypothetical TCC is shown in Figure 1. Assuming that some items from the test represented by this TCC are to be administered in some adaptive manner (e.g., MISS) and that a level of θ is to be determined that corresponds to, say, 70% correct performance on the entire test, it may be done using the following steps:

1. Draw a horizontal line (Line A in Figure 1) from the .7 mark on the vertical (expected proportion correct, or P) axis of the TCC figure to the TCC.

2. Drop a vertical line (Line B) from the point of intersection of the TCC and the horizontal line drawn in Step 1 to the horizontal (achievement level, or θ) axis. This point (θ_m) on the achievement level axis is designated the mastery level in terms of the achievement (θ) metric.

3. The mastery level specified in Step 2 above may now be used to make mastery decisions in place of the .7 mastery level originally specified using any subset of items from the original test, provided that individuals' item responses are scored with a method that will put the θ estimate on the same metric as the TCC. Any ICC-based scoring procedure (e.g., Bejar & Weiss, 1979) will result in a θ estimate that will be on the correct metric. This procedure allows the transformation of any desired proportion correct mastery level to the θ metric. Once this transformation is made, ICC theory and its technology may be used to increase the efficiency of present mastery testing techniques.

Figure 1
Hypothetical Test Characteristic Curve Illustrating Conversion from the Proportion Correct Metric to the Achievement Metric



Making the mastery decision using Bayesian confidence intervals. Although any achievement level estimate of any subset of the items from a test obtained using ICC-based scoring will be on the same metric as the TCC for the original test, two different subsets of items may result in θ estimates that are not equally informative. For example, if one test consisted of many items and the other used only a few items, the longer test would probably yield a more precise θ estimate, provided that the items in the two tests had similar ICCs. Thus, ICC-based θ estimates that are on the same metric are comparable except for their differential precision. Comparisons of ICC-based θ estimates should therefore be based on confidence interval estimates instead of the raw achievement level point estimates.

For this reason, the AMT strategy makes mastery decisions with the use of Bayesian confidence intervals. Specifically, after each item is selected and administered to an individual--for this application MISS is used to choose the appropriate item at each point in the test-- a point estimator of the individual's achievement level ($\hat{\theta}$) may be determined using Owen's Bayesian scoring algorithm, using information gained from all items administered previously. Given this point estimate and the corresponding variance estimate, also obtained using Owen's procedure, a Bayesian confidence interval may be defined such that

$$\hat{\theta}_i - 1.96(\sigma^2_i)^{\frac{1}{2}} \leq \theta \leq \hat{\theta}_i + 1.96(\sigma^2_i)^{\frac{1}{2}} \qquad \text{with } p = .95 \qquad [10]$$

where

$\hat{\theta}_i$ = the Bayesian point estimate of achievement level, calculated following item i;

σ^2_i = the Bayesian posterior variance following item i; and

θ = the true achievement level.

Equation 10 may be interpreted as meaning that the probability is .95 that the true value of the achievement level parameter, θ, is within the bounds of the confidence inteval. It might also be said that there was 95% confidence that the true parameter value lies within the confidence interval.

After this confidence interval has been generated, it is a simple matter to determine whether or not θ_m, the achievement level earlier designated as the mastery level on the achievement metric, falls outside the limits of the confidence interval. If it does not, the testing procedure administers another item to the individual and recalculates the confidence interval. This procedure continues until, after some item has been administered, the confidence interval calculated will not include θ_m, the mastery level on the achievement continuum. At this time the testing procedure terminates and a mastery decision is made. If the lower limit of the confidence interval falls above the specified mastery level, θ_m, the individual is declared a master; if the upper limit of the confidence interval falls below θ_m, the individual is declared a nonmaster. Given a finite item pool size, however, the testing procedure may exhaust the pool before a decision can be made in this manner. It is possible to make a decision concerning mastery for any of these individuals based on whether the Bayesian point estimate of their achievement level ($\hat{\theta}$) is above or below the specified mastery level, θ_m. These decisions, however, cannot be made with the same degree of confidence as those made with confidence intervals that do not contain the mastery level.

Wald's SPRT versus ICC-Based AMT Procedure

The two mastery testing strategies described above differ in a number of characteristics. The most salient of these differences are as follows:

1. Treatment of the items in the domain.
2. Treatment of the uncertainty of decisions.
3. Treatment of the mastery cutoff.
4. Treatment of the achievement metric.

Treatment of items. The SPRT in the simple form outlined above, treats all of the items in the mastery test as if they were perfect replicates of each other. Thus, an individual's response to a particular item is viewed solely as a probabilistic function of the individual's true mastery status. This assumption is most appropriate in the production setting in which Wald originally designed his procedure; each light bulb can be expected to be like every other light bulb. This assumption may be less tenable in the mastery testing situation, where an individual's responses to test items may vary as a function of differential characteristics of the items themselves, as well as his/her mastery status.

The AMT procedure assumes that if items differ, their individual characteristics may be described by a logistic ogive that varies as a function of the item's power to discriminate among individuals with different achievement levels (a), the item's difficulty (b), and the ease with which an individual may answer the item correctly with no knowledge of the subject mattery (c). This assumption concerning the operating characteristics of the items is less restrictive than the assumption made in the SPRT procedure described above; but to the extent that the items do not conform to the logistic form specified, the assumption might still restrict the efficiency of the AMT procedure.

Both mastery testing procedures, therefore, postulate some systematic similarities among the test items. To the extent that one of the postulations is closer to the actual state of the world than the other, it might be expected that the corresponding procedure would perform more efficiently. Thus, the characteristics of the item pool to be used for mastery testing yields the first point at which it might be decided which of the two models is more appropriate for use in a given situation.

Treatment of uncertainty. The SPRT makes use of traditional hypothesis testing methods to determine the point at which an individual's item responses are sufficient evidence for making a decision concerning his/her mastery status. Here "sufficient" is defined in terms of the α and β error rates that one is willing to accept across all the students tested. α and β may be set independently to reflect the educator's concerns over the relative costs of the two error types.

The AMT procedure uses a symmetric Bayesian confidence interval to make the mastery decision. This functionally sets α equal to β and, by doing so, implies equal costs for the two error types. To the extent that the costs of the two error types are not equal, the SPRT provides the educator with more flexibility than the AMT procedure, as currently operationalized.

Treatment of mastery level. The SPRT uses an uncertainty region, rather than a single mastery level, to define the mastery and nonmastery regions. The specification of this uncertainty region is based on a decision by the educator concerning the range that appropriately reflects uncertainty as to whether the student's performance is actually the performance of a master or a nonmaster. By contrast, the AMT procedure defines a single mastery level and determines whether an individual is significantly above or below the mastery level using a Bayesian confidence interval.

This difference between the two testing procedures renders tentative any comparison that might be made. The performance of the SPRT procedure will vary widely as a function of the uncertainty band chosen. For the AMT technique this uncertainty is not directly taken into account. Any comparison between the two techniques is conditional upon the width and absolute bounds of the uncertainty region.

Treatment of the θ metric. The decisions made by the SPRT are dependent on the percentage of items that are correctly answered for any specific test length. Thus, the metric of achievement assumed in this procedure is the proportion-correct metric. The AMT procedure assumes, due to the differential properties of the items in the item pool, that there is a nonlinear transformation of the proportion-correct metric, which more accurately represents the achievement of the individuals taking the test. This latent continuum serves as the achievement metric for the AMT procedure.

This difference in the achievement metric again renders comparisons between the two procedures somewhat difficult, since the "true" achievement levels of individuals must be postulated to fit one of these metrics. Any differences noted in the performance of the two procedures may be due to this difference in the achievement metrics assumed.

EMPIRICAL COMPARISON OF THE SPRT AND AMT PROCEDURES

To delineate circumstances in which one of the mastery testing procedures might have an advantage over the other, monte carlo simulation was used to compare the two testing procedures under several conditions.

Method

The method used to compare the two variable-length mastery testing procedures to one another, as well as to a conventional (fixed length) testing procedure, consisted of five basic steps:

1. Three item pools were generated in which the items differed from one another to different degrees.

2. Item responses were generated for 500 simulated subjects (simulees) for each of the items in the three item pools.

3. Conventional tests of three different lengths were drawn from the larg-

er item pools; these conventional tests served as item pools from which the SPRT and AMT procedures drew items.

4. The AMT and SPRT procedures were simulated for each of the three different item pool types and the three conventional test lengths.

5. Comparisons were drawn among the three types of tests (AMT, SPRT, conventional) concerning the degree of correspondence between the decisions made by the three test types and the true mastery status. Further comparisons were made based on the average test length that each test type required to reach its decisions.

Item Pool Generation

Three 100-item pools were generated to reflect different types of pools that might be used in a mastery test.

Uniform pool. The uniform pool consisted of 100 items that were perfect replications of one another. Each item had the same discrimination ($a = 1.00$), difficulty ($b = 0.00$), and guessing probability ($c = .20$). This pool was designed to correspond to the SPRT procedure's assumption that all items in the test are similar.

b-variable pool. The b-variable pool varied from the uniform pool only in that the items had a range of difficulty levels. Eleven values of b were assigned to an approximately equal number of items in the pool. The values of b chosen were -2.50, -2.00, -1.50, -1.00, -0.50, 0.00, 0.50, 1.00, 1.50, 2.00, and 2.50. Nine items at each level of difficulty were used in this pool, along with an additional item with $b = 0.00$ to bring the pool to 100 items.

a-, b-, and c-variable pool. The a-, b-, and c-variable pool differed from the b-variable pool in that the discriminations and guessing levels of the items were allowed to spread across a range of values. The a values used were .50, 1.00, 1.50, and 2.00. The c values used were .10, .20, and .30. All a and c values were approximately equally represented. The parameter estimates were arranged such that each level of difficulty was represented by items that had approximately the same average a level and the same average c level (i.e., the pool was approximately rectangular).

Item Response Generation

Achievement levels for 500 simulees were drawn from a normal distribution with a mean of zero and a standard deviation of one. Item responses for each of these simulees were then generated for each item in each of the three item pools using the 3-parameter logistic ICC model. That is, knowing the θ level of the simulee and the parameters of the item in question, the probability of a correct response was calculated. A random number was then drawn from a uniform distribution ranging from zero to one. If this number was lower than the probability of a correct response, the simulee was given a correct response to the item. If the number was higher than the correct response probability, the simulee was given an incorrect response.

Thus, in this study, the achievement metric and the item response generator correspond closely to the model assumed by the AMT procedure. The "true" mastery level for each simulee was determined by comparing the θ levels used to generate the item responses with the proportion correct mastery level expressed on the θ metric.

Conventional Tests

Conventional tests of three different lengths (10, 25, and 50 items) were drawn at random from each of the three item pools, with the stipulation that the shortest conventional test served as the first portion of the next longer conventional test and that this test in turn served as the first portion of the longest conventional test. These nine conventional tests served as subpools from which the AMT and SPRT procedures drew items during the simulations. This random sampling from a larger domain of items was designed to correspond to the traditional mastery testing paradigm and to the random sampling model underlying the SPRT.

Simulation of the Testing Strategies

Using the item response data for the 500 individuals and the item parameters available for each of the items (for the AMT procedure), the three testing strategies (AMT, SPRT, conventional) were employed to make mastery decisions for each individual. Each testing procedure was used with each of the nine subpools.

Conventional test. The conventional test assumed a mastery criterion of 60% correct responses. After all of the items in the conventional test were administered, if the individual answered 60% or more items correctly, the individual was declared a master. If the individual's score was less than 60% correct, the individual was declared a nonmaster.

SPRT procedure. For the SPRT procedure the limits of the uncertainty region were set at proportion-correct values of .50 and .70. Values of α and β were each set to .10. For individuals for whom no decision was made by the Wald procedure before the item pool was exhausted, the mastery decision was made by the conventional procedure, using a mastery proportion of .60.

AMT. For the AMT procedure the mastery levels in each of the 100-item pools corresponding to 60% correct were designed to be equal to $\theta = 0.00$. This mastery level was used with each of the smaller item pools, even though they had not been designed to result in a mastery level of $\theta = 0.00$. This procedure added some sampling error to the AMT procedure, to more appropriately reflect the error that is inherent when using estimated item parameters to determine the mastery level. For the AMT Bayesian scoring procedure, each individual was assumed to have a prior mean of 0.00 and a prior variance of 1.00.

Comparison among Testing Procedures

For each of the three testing procedures (AMT, SPRT, conventional), the value of the procedure may be judged by the average length of the test required to make the mastery decision and by how well the decisions that are made reflect the true state of nature. Specifically, the AMT and SPRT procedures were compared in terms of the average reduction in the length of the test required to make mastery decisions across the entire group of individuals. Further, all three procedures were compared in terms of how well the decisions they made corresponded with the true mastery status of the individuals.

Comparisons within each testing procedure concerning the average test length and the correspondence of decisions with true mastery status were made across all nine combinations of test lengths and item pool types.

RESULTS

Test Length

Table 1 shows the mean test length required by each of the testing procedures to make a decision concerning the mastery status of the simulees in the test group.

Table 1
Mean Number of Items Administered to Each Simulee
for Three Mastery Testing Strategies Using Each Type
of Item Pool, at Three Maximum Test Lengths

Item Pool and Testing Strategy	Maximum Test Length 10	25	50
Uniform Pool			
Conventional	10.00	25.00	50.00
AMT	9.03	15.99	23.00
SPRT	8.75	13.12	15.39
b-Variable Pool			
Conventional	10.00	25.00	50.00
AMT	9.43	18.09	27.17
SPRT	9.62	16.79	21.41
a-, b-, and c-Variable Pool			
Conventional	10.00	25.00	50.00
AMT	8.73	16.35	23.39
SPRT	8.62	13.42	15.70

Uniform pool. As can be seen from Table 1, the AMT procedure resulted in some test length reduction for each maximum test length (MTL), with the reduction in test length increasing as the MTL increased. For the 10-item MTL, the percentage by which the conventional test length was reduced was 9.7%; for the 25-item MTL the reduction was 36%; and for the 50-item MTL the observed reduction was 54%.

For the SPRT procedure, again, increasing test length reduction was noted as MTL increased; and some reduction was noted at each level of MTL. For the 10-item MTL, the reduction observed was 12%. The 25-item MTL resulted in a 48% reduction. For the 50-item MTL the reduction was 69%. At all MTL levels the SPRT procedure resulted in a greater reduction of test length than the AMT procedure.

b-variable pool. For the pool in which the difficulty levels of the items differed, the data in Table 1 show the same trends that were noted for the uniform pool. The AMT procedure reduced the test length at each MTL, and the reduction increased with the MTL level. For the 10-item, 25-item, and 50-item MTL levels, the AMT procedure reduced test length by 6%, 28%, and 46%, respectively.

The SPRT procedure also reduced test length at each MTL level, with larger reductions for the longer MTL levels. At the 10-item, 25-item, and 50-item MTL levels the test length reductions observed were 4%, 33%, and 57%, respectively.

For this pool the AMT procedure resulted in slightly greater reduction in test length at the 10-item MTL level, whereas the SPRT procedure resulted in greater test length reductions for the longer MTL levels. Across all MTL levels, both procedures reduced test length somewhat less for this item pool than for the uniform item pool.

a-, b-, and c-variable pool. Table 1 shows that when the AMT procedure was used with this item pool, test length was again reduced at each MTL and this reduction was greater for the longer MTL levels. For the 10-item, 25-item, and 50-item MTL levels, the observed reductions in test length were 13%, 35%, and 53%, respectively.

For the SPRT procedure with this item pool, test length reduction was once more observed, with an increasing reduction as the MTL increased. The reductions noted were 14%, 46%, and 69% for the 10-item, 25-item, and 50-item MTL levels.

For this item pool the SPRT procedure terminated using a smaller average number of items for each MTL. Further, the degree of test length reduction in this pool for both procedures, at all MTL levels, was quite similar to that observed for the uniform item pool.

Correspondence with True Mastery Status

For each of the simulees in the sample, the true θ level was known: It was the level that was used to generate the item responses. Given this, it was known whether the individual's θ level was actually above or below the prespecified mastery level on the achievement metric ($\theta = 0.00$). Phi correlations between true mastery status and the mastery state determined by each of the three testing procedures for each MTL level and pool type are shown in Table 2.

Uniform pool. For the uniform pool one major trend was observed. For each testing procedure an increase in the MTL level was accompanied by an increase in the correlation between the true and estimated mastery states. (These correlations may be referred to as correspondence coefficients.)

Table 2
Phi Correlations Between Observed Mastery State and True Mastery State for Each Mastery Testing Strategy, Using Each Type of Item Pool, at Three Maximum Test Lengths

Item Pool and Testing Strategy	Maximum Test Length 10	25	50
Uniform Pool			
Conventional	.771	.837	.875
AMT	.775	.840	.871
SPRT	.771	.837	.867
b-Variable Pool			
Conventional	.541	.667	.783
AMT	.615	.715	.828
SPRT	.541	.656	.704
a-, b-, and c-Variable Pool			
Conventional	.290	.670	.735
AMT	.470	.733	.787
SPRT	.290	.592	.571

In addition to this major trend, it was observed that for the 10-item and 25-item MTL levels, the AMT procedure produced the highest correspondence coefficient observed (r = .775 and .840, respectively). For the 50-item MTL level the conventional procedure resulted in the highest correspondence (r = .871).

It should be noted that the differences in correspondence between any two MTL levels within any testing procedure (the smallest was .03, between the 25-item and 50-item MTL levels for the SPRT procedure) were much larger than the largest difference noted between any two testing procedures within a single MTL level (.008, for the conventional and SPRT procedures in the 50-item MTL level).

b-variable pool. The same major trend that was found for the uniform pool was again observed in the b-variable pool. Each testing strategy resulted in higher correspondence as the MTL level increased. For each MTL level, the AMT procedure resulted in the highest correspondence coefficients. The conventional procedure resulted in the next highest correspondence level for all three MTL levels (tied with the SPRT procedure at the 10-item MTL level).

Differences in correspondence coefficients observed between testing strategies within an MTL level were larger in this pool than in the uniform pool but were still somewhat smaller than the differences noted between MTL levels, on the average. It was also noted that each correspondence level observed was lower for this pool than for the uniform pool across all MTL levels and testing procedures.

a-, b-, and c-variable pool. The same trend of increasing correspondence with increasing MTL level was again noted for the conventional and AMT procedures. For the SPRT procedure the correspondence peaked at $\underline{r} = .592$ at the 25-item MTL level and dropped to .571 at the 50-item MTL level.

The AMT procedure produced the highest correspondence for all three MTL levels. The conventional procedure resulted in the next highest level of performance at all MTL levels (again tied with the SPRT procedure at the 10-item MTL level).

Once again, the average difference in correspondence was much greater between MTL levels within testing strategies than between two testing strategies within a single MTL level. Further, on the average, the correspondence coefficients for this pool were lower than for either of the other pools, with rather large decreases at the 10-item MTL level, particularly for the conventional and SPRT strategies.

Frequency and Type of Errors

To further compare the performance of the three mastery testing strategies the frequency with which each procedure made incorrect decisions (false mastery, false nonmastery) was examined; the percentage of decision errors made by each of the testing strategies with each of the item pools at each MTL is shown in Table 3. This table shows the frequency with which each of the testing procedures made false mastery and false nonmastery decisions in each of the testing conditions. It may be noted that the "Total" column in Table 3 reproduces the information already reported from the correlational analysis, but in a different manner. For each situation in which a high correlation was noted, a correspondingly low total error rate is noted in Table 3, as expected.

Uniform pool. For the uniform pool each of the testing strategies resulted in the same general pattern of errors across MTL levels. Each procedure resulted in more false nonmastery decisions than false mastery decisions at all MTL levels. Each procedure also resulted in fewer errors of each type with increased MTL. The difference in the frequencies of false mastery and false nonmastery decisions was smaller with larger MTL levels for all procedures. The differences among the procedures in terms of the types of false decisions made were minimal.

b-variable pool. For this item pool the patterns of errors made by the different testing strategies were less regular than in the uniform pool. The conventional and SPRT procedures produced more false mastery than false nonmastery decisions at all MTL levels. The AMT procedure produced more false mastery than false nonmastery decisions at the 10-item MTL level but produced more false nonmastery than false mastery decisions at the two higher MTL levels. For the AMT procedure the discrepancy in the frequencies of the two types of errors was smaller than for the other two procedures at all three MTL levels and was quite small (less than 2%) at the two higher MTL levels. For the conventional procedure the difference in the frequencies of the two types of errors was quite small at the highest MTL level; but for the SPRT procedure, a fairly large discrepancy between the two error rates (20% to 80%) was observed at each MTL level.

Table 3
Percentage of Incorrect Decisions by Type of Error Made by Each Testing Strategy, Using Each Type of Item Pool, at Three Maximum Test Lengths

	Maximum Test Length								
	10			25			50		
Item Pool and Test	False Mastery	False Non-Mastery	Total	False Mastery	False Non-Mastery	Total	False Mastery	False Non-Mastery	Total
Uniform Pool									
Conventional	3.6	8.0	11.6	2.6	5.6	8.2	2.8	3.4	6.2
AMT	3.6	7.8	11.4	3.0	5.0	8.0	3.0	3.4	6.4
SPRT	3.6	8.0	11.6	2.6	5.6	8.2	3.2	3.4	6.6
b-Variable Pool									
Conventional	22.4	2.2	24.6	13.4	3.6	17.0	6.4	4.4	10.8
AMT	12.2	7.0	19.2	6.6	7.6	14.2	3.4	5.2	8.6
SPRT	22.4	2.2	24.6	14.2	3.4	17.6	11.4	3.6	15.0
a-, *b*-, and *c* Variable Pool									
Conventional	0.0	44.6	44.6	2.6	15.2	17.8	7.4	5.8	13.2
AMT	8.0	19.4	27.4	5.2	8.2	13.4	5.0	5.6	10.6
SPRT	0.0	44.6	44.6	2.0	21.0	23.0	3.8	19.4	23.2

In all testing conditions but one (AMT with a 25-item MTL), the use of the *b*-variable item pool resulted in higher discrepencies between the two observed error rates (as well as higher absolute error rates) than when the uniform pool was used.

a-, b-, and c-variable pool. For this item pool, each of the testing procedures resulted in higher frequencies of false nonmastery decisions than false mastery decisions for the 10-item and 25-item MTL levels. For the 50-item MTL level the conventional procedure resulted in a higher frequency of false mastery decisions, but the AMT and SPRT procedures still resulted in higher percentages of false nonmastery decisions. As with the *b*-variable item pool, the AMT procedure used with this item pool resulted in smaller differences in the frequencies of the two error types than either of the other testing procedures at each MTL level. For the 50-item MTL level the AMT procedure produced a very small difference in the two error rates (.6%). The conventional procedure also produced a small difference in the two error rates for the 50-item MTL level (1.6%). The SPRT procedure resulted in the highest difference between the two error rates at all MTL levels (tied with the conventional procedure at the 10-item MTL level).

One interesting result was observed when the errors made with the *b*-variable item pool were compared with those made using the *a*-, *b*-, and *c*-variable item pool. For the *b*-variable pool each of the testing procedures was more likely to make false mastery decisions than false nonmastery decisions. This tendency was reversed for the *a*-, *b*-, and *c*-variable item pool, where each of the procedures made more false nonmastery decisions than false mastery decisions. These trends were most noticeable for each of the testing procedures at the 10-item MTL level, and most noticeable for the SPRT procedure across all MTL levels. It is probable that these trends were artifacts of the random sampling of items used to create the conventional tests, since the shorter conventional tests would be less representative of the item domain due to the small sample of items taken. The results obtained here would be explained by a very easy 10-item conventional test being drawn from the *b*-variable pool and a very difficult 10-item test being drawn from the *a*-, *b*-, and *c*-variable pool. In fact, the mean *b*-value for the 10-item conventional test drawn for the *b*-variable pool was -.80; for the *a*-, *b*-, and *c*-variable pool, it was 1.25. This would also explain the observation that the SPRT procedure most clearly showed these trends, since the SPRT procedure used shorter test lengths, on the average, than the other two procedures to make its final decisions and therefore was most prone to small-sample artifacts.

DISCUSSION AND CONCLUSIONS

Several trends were noted in the data concerning the performance of the three testing strategies in the three different item pools. In every instance the AMT and SPRT procedures produced reductions in the mean test length required to make mastery decisions. This reduction increased with the MTL level in each circumstance. The AMT procedure resulted in reductions of 6% to 54% from the length of the conventional test. The SPRT procedure resulted in reductions of 4% to 69%. On the average, the SPRT procedure required fewer items to make the mastery decision.

The correspondence between the estimated mastery status and the true mastery status systematically increased with MTL for all testing procedures in each item pool. The correspondence fairly systematically decreased from the uniform pool, to the b-variable pool, to the a-, b-, and c-variable pool. The AMT procedure resulted in the highest level of correspondence in all circumstances but one (the conventional test performed best for the 50-item MTL with the uniform pool). On the average, though, the differences between different MTL levels were more pronounced than differences between testing procedures. Further, the type of item pool used had pronounced effects on the correspondence obtained.

The AMT procedure resulted in the most even frequencies in the types of decision errors made across most MTL levels and item pools. This was desirable, since both error types were assumed to have the same relative cost. Further, it was noted that the SPRT procedure was most susceptible to small-sample artifacts, resulting in an imbalance in the frequencies with which the two types of errors were made.

To prescribe the best testing strategy of those described here requires specification of priorities and conditionals. If a uniform item pool is assumed, the SPRT procedure required the fewest items while resulting in decisions having correspondence coefficients that were quite comparable to the other two procedures. If, however, the item pool includes items with variable a, b, and c parameters, the SPRT procedure may result in the shortest tests, but the AMT procedure will make more accurate classifications. These factors must be considered before any decision is made as to which procedure is "best."

It should also be noted that this simulation was based on the assumption that the latent achievement metric, rather than the propertion-correct metric, was the correct metric; and to the extent that the proportion-correct metric is the correct metric, the findings of this study are less relevant. In addition, several variations on the SPRT procedure and the AMT procedure that were not examined in this study are possible; thus, additional research is necessary before firm conclusions can be drawn concerning the utility of adaptive mastery testing strategies.

REFERENCES

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A047495)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Brown, J. M., & Weiss, D. J. An adaptive testing strategy for achievement test batteries (Research Report 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1977. (NTIS No. AD A046062)

Ferguson, R. L. Computer-assisted criterion-referenced measurement (Working Paper No. 41). University of Pittsburgh, Learning and Research Development Center, 1969. (ERIC Document Reproduction No. ED 037 089)

Ferguson, R. L. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. (Doctoral dissertation, University of Pittsburgh, 1969) Dissertation Abstracts International, 1970, 30, 3856A. (University Microfilms No. 70-4530).

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.), Psychological principles in system development. Chicago: Holt, Rinehart, & Winston, 1962.

Kingsbury, G. G., & Weiss, D. J. An adaptive testing strategy for mastery decisions (Research Report 79-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1979. (a)

Kingsbury, G. G., & Weiss, D. J. Relationships among achievement level estimates from three item characteristic curve scoring methods (Research Report 79-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, April 1979. (NTIS No. AD A069815) (b)

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

McBride, J. R., & Weiss, D. J. Some properties of a Bayesian adaptive ability testing strategy (Research Report 76-1). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, March 1976. (NTIS No. AD A022964)

Nitko, A., & Hsu, T. C. Using domain referenced tests for student placement, diagnosis, and attainment in a system of adaptive individualized instruction. Educational Technology, 1974, 14, 48-53.

Owen, R. J. A Bayesian approach to tailored testing (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service, 1969.

Wald, A. Sequential analysis. New York: Wiley, 1947.

ACKNOWLEDGMENTS

This research was supported by funds from Air Force Office of Scientific Research, Army Research Institute, Defense Advanced Research Projects Agency and Office of Naval Research, under Contract N00014-79-C-0172 NR 150-433 with the Personnel and Training Research Programs, Office of Naval Research.

DISCUSSION: SESSION 3

MELVIN NOVICK
UNIVERSITY OF IOWA

I shall discuss some general methodological issues that bear on the papers by Reckase, Kalisch, and Kingsbury and Weiss and also on previous papers presented, integrating into the discussion relevant points that have been made by Lord, Wainer, Samejima, Lumsden, and others.

The results that have been obtained in these papers are contradictory. There seems to be difficulty deciding whether or not adaptive testing is worthwhile with a Bayesian approach--which is related to the kinds of models that have been adopted and the kinds of statistical analysis that are being performed. Lord made an important comment about the metric in which a least squares analysis is performed; and although the suggestion he made in that context was very good, it opens up the question, which is the correct metric? Wainer's comments about robustness are also very important; indeed, some of the problems that we have had have resulted from allowing a few outliers to mar the analyses. An important part of my discussion will also bear on his comment, "Let's look at the ends, because it doesn't matter what's going on in the middle." Samejima's comments about dimensionality are crucial; and Lumsden's comment about the importance of choosing the statistical analysis for the particular decision at hand is central to my discussion.

I am absolutely delighted to see that everyone is using Bayesian methods: It is a dream come true. The realization of the dream, however, remains imprecise. Although Bayesian procedures are being used, the analyses are not all Bayesian which is part of the problem I hope to correct.

A brief discussion is in order about the development of pre-Bayes statistical theory and its application in a Bayesian decision theoretic context. First was Gauss's work on least squares, which led to a certain mean value as an estimate; this was followed by the Gauss-Markov theorem, which tied least squares with the normal distribution. At about the same time, La Place was working with absolute error loss, which is typically a better loss function than squared error, and in my judgment La Place deserves more credit than he has been given. Once the question, how to obtain an estimator, has been posed and considered in terms of the appropriate loss function, a whole new set of problems arises.

Even though absolute error loss may be better than squared-error loss, in some of the applications this places too much weight on those large discrepancies. In terms of mastery testing, for example, it does not matter if the person is three standard deviations from the criterion or four. Certainly, as Wainer has said, we do not want the analysis to be affected very much by that, particularly when it is recognized that the distributions are not normal but that there are all kinds of outliers and unusual data values. This is not a minor point. It affects all the analyses that are being done. A very careful look must be taken at the loss function in deciding whether the decision rule or even an estimator is any good. In my judgment, none of the loss functions that have been talked about at this conference are acceptable.

I did use threshold loss in papers published several years ago; but at that time, Bayesian methods with threshold loss were better than classical methods. Now there are better methods, and recent papers discussing more general loss or utility functions provide much more acceptable methods. In these papers the normal ogive is used as a utility function. This is a clear improvement over threshold utility. However, there is a Stage 3 in which a cumulative data distribution may be used--perhaps some other ogival forms--as a utility function.

One of the techniques that was used in a paper in an earlier session was to ascertain how adaptive testing improves reliability and squared-error loss. The difference between looking at a reliability and looking at a squared-error loss is that reliability forgets about any kind of bias. However, squared-error is actually irrelevant to a context in which a mastery decision or a selection is being made. This does not mean that I repudiate either classical test theory, which is built largely on mean-squared-error, or latent trait theory. Those methods are useful in certain contexts, e.g., when developing a test that is going to be used for a wide range of purposes, when interest is in discrimination across the whole range of ability and some overall measure is needed, and for the SAT and ACT tests. These methods are much less useful in the context in which there is a question of mastery or selection and one has a fair idea where that selection is going to be. Then, it is desirable to peak the tests at roughly that point, but it is also desirable to use a loss function, or better yet, a utility function that focuses on that point. Therefore, looking at questions of reliability and squared-error loss does not really address the question of the efficacy of the procedure in any real way.

On a related issue, there has been discussion on using Bayesian modal estimates or maximum likelihood estimates, which are, of course, also Bayesian modal estimates assuming a uniform prior distribution on a particular parameterization. These are appropriate only in terms of a zero-one loss function, a most unrealistic loss function in this context. Therefore, the analyses based on maximum likelihood or a Bayesian modal estimator may be unrealistic.

The dimensionality issue is crucial. Something like the reliability-validity paradox may, in fact, be occurring here, as Samejima suggests. It would not surprise me at all if we are dealing with a test that is multidimensional and a criterion that is almost certanly multidimensional. If this is true, and if a dominant trait is focused on, we may be building up reliability and not measuring the other traits that are essential in prediction. Thus, validity will suffer. The answer to this is probably to study the predictor and the criterion carefully and to define the factors or traits and see that each one is measured carefully.

Next, if least squares is to be used, which I do not really advocate, there is the question of what metric to do it in. Should it be done in a latent variable metric? This causes problems because computations sometimes do not converge. Should it be done in the true score metric, which is tighter? Although I do not know the answer to that question at present, the question should not be ignored.

The questions that need to be considered are (1) How much efficiency is being obtained? (2) Where is the efficiency being sought? and (3) What is the appropriate measure of efficiency? If a procedure is being designed to assist in selection near the top of the distribution or at some other criterion point, it really does not matter whether or not better estimation is being obtained away from that point. It is totally irrelevant to state that there is only a 5% increase in efficiency overall. A 50% increase could be obtained where needed, still averaging out to 5% overall. That would not be bad. This is a question of how the gains are computed, which, again, may be related to the question of robustness. We may be doing terrible things with some outlier; but if a testee is completely off the scale, perhaps it does not really matter, because a large error will not affect the decision.

The Owen procedure is a good Bayesian procedure: It does make some assumptions. Even though some of the assumptions that it makes may not be terribly well satisfied for the first one-half dozen items, improvements are possible, but that is not important. If any reasonable Bayesian procedure is used, a great deal will be gained from the Bayesian allocation. If a person is seated at a terminal, it may not be very significant whether he/she takes 5 items or 6 items. Thus, I am not so sure that the emphasis on variable stopping is important. Some rules could probably be worked out that, by and large, would provide good results if all testees were given a Bayesian allocated test of specified length and the decision were made at that point. The advantage would be that most of the inaccuracies in the approximations of the Owen procedure would be eliminated. If, indeed, the saving of one item, on the average, has a high pay off, then presumably someone would be willing to make a large investment to obtain the needed refinements.

Now, I should like to treat some specifics of the Reckase and Kingsbury and Weiss papers. In each paper there is an emphasis on the Wald Sequential Probability Ratio Test (SPRT). The original application of this method was that there was a production process in control with a certain error rate that was tolerable. The concern was that something had happened that seriously degraded production quality and it was desirable to identify the problem very quickly. Therefore, it was very reasonable to take a certain point hypothesis, a 3% error rate, with the recognition that if the process was not in proper working order, that error rate was going to go up to 10% and therefore the alternative hypothesis of 10% should be used. That paradigm is not correct in the context of adaptive testing. What the SPRT formulation gives is utility functions with three levels corresponding to the false positive, false negative, and indifference zones. In fact, an appropriate utility function would be continuous and not abrupt in change of magnitude of the first derivative (see Novick & Lindley, 1978). This is very important because it has a very substantial effect on the analysis, both in terms of the number of observations needed and in terms of the decision rule to be adopted.

A minor technical point is that one simply cannot look ahead one step, computing the cost of taking an observation and comparing this with the expected gain, and then stopping when it is discovered that the expected gain does not exceed the cost of the observation. In fact, all possible sample sizes would have to be investigated to make sure that none yielded an expected gain. I am not, however, arguing for this complication; indeed, I am arguing for a simplification to fixed sample sizes.

Finally, although there are a half a dozen other examples within an epsilon of the one I selected to discuss, Kingsbury and Weiss's paper presents the most simple and striking example of doing Bayesian analysis without a saturation of understanding Bayesian theory. The idea of looking at the Bayesian confidence interval, or as I would prefer to call it, the Bayesian credibility interval, and then stopping when that Bayesian interval no longer included a particular point is perfectly reasonable. In the context of mastery decision making, however, I cannot understand why a two-sided interval was computed. It makes no sense at all from any kind of Bayesian logic. Any consideration of a concept of utility or loss must lead to a one-sided interval. That struck me as being the most glaring failure to bring decision theory to bear on what is being done. If pressed, however, I could find a half a dozen more examples; and that, I think, is discouraging.

REFERENCES

Novick, M. R., & Lindley, D. The use of more realistic utility functions in educational applications. Journal of Educational Statistics, 1978, 15, 81-91.

SESSION 4:
ESTIMATING RESPONSE FUNCTIONS WITHOUT ASSUMING A PARAMETRIC MODEL

CONSTANT INFORMATION MODEL ON THE DICHOTOMOUS RESPONSE LEVEL

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE

DISCUSSION

ROBERT TSUTAKAWA
UNIVERSITY OF MISSOURI--
COLUMBIA

CONSTANT INFORMATION MODEL ON THE DICHOTOMOUS RESPONSE LEVEL

FUMIKO SAMEJIMA
UNIVERSITY OF TENNESSEE

Generally speaking, the fundamental role of the mathematical model in psychology is to simulate psychological reality following some sound rationale, using well-defined parameters. The normal ogive model in latent trait theory, for example, is one of such mathematical models. Another role of the mathematical model may be to provide some mathematical convenience, just as the logistic model does in its relationship with the normal ogive model.

Mathematical models of a third type, which have their specific usefulness in the context of comprehensive theories and methods, can be conceived. The direct simulation of psychological reality is less important for this type of model than it is for the first two types of models. The Constant Information Model belongs to this new type; its role is to be of assistance in the developmental stages of theories and methods, rather than to simulate psychological reality directly.

The Constant Information Model (Samejima, 1979) is a new model on the dichotomous response level (Samejima, 1972). This model provides a constant item information for a finite interval of a latent trait. Although the usefulness of the model has not yet been fully investigated, an effective use of the model has been found in the process of estimating the operating characteristics of item response categories (Samejima, 1977b, 1977d, 1978a, 1978b, 1978c, 1978d, 1978e, 1978f).

Let θ be ability, or latent trait, which is assumed to be unidimensional. Let $\underline{g}$ (= 1, 2,..., $\underline{n}$) be an item and $\underline{x}_g$ (= 0, 1, 2,..., $\underline{m}_g$) denote an item response category, or item score. The operating characteristic, $P_{x_g}(\theta)$, of the item score $\underline{x}_g$ is defined by

$$P_{x_g}(\theta) = \text{prob.}[x_g|\theta] \quad . \qquad [1]$$

Let V be a response pattern, such that

$$V = (x_1, x_2, \ldots, x_g, \ldots, x_n)' \ , \qquad [2]$$

and $P_V(\theta)$ be its operating characteristic. Because of the assumption of local independence,

$$P_V(\theta) = \prod_{x_g \varepsilon V} P_{x_g}(\theta) \qquad [3]$$

can be written. The *item response information function* (Samejima, 1969), $I_{x_g}(\theta)$, is defined by

$$I_{x_g}(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_{x_g}(\theta) \qquad [4]$$

and the *item information function*, $I_g(\theta)$, is the conditional expectation of the item response information function, given θ, so that

$$I_g(\theta) = E[I_{x_g}(\theta)|\theta] = \sum_{x_g=0}^{m_g} I_{x_g}(\theta)\, P_{x_g}(\theta)$$

$$= \sum_{x_g=0}^{m_g} [\frac{\partial}{\partial\theta} P_{x_g}(\theta)]^2\, [P_{x_g}(\theta)]^{-1} \qquad [5]$$

(cf. Samejima, 1969, chap. 6). The *response pattern information function*, $I_V(\theta)$, can be written for a specified response pattern V such that

$$I_V(\theta) = -\frac{\partial^2}{\partial\theta^2} \log P_V(\theta) = \sum_{x_g \varepsilon V} I_{x_g}(\theta) \quad , \qquad [6]$$

and the *test information function*, $I(\theta)$, is defined as the conditional expectation, given θ, of the response pattern information function, which can be written

$$I(\theta) = \sum_V I_V(\theta)\, P_V(\theta) = \sum_{g=1}^{n} I_g(\theta) \ . \qquad [7]$$

When item *g* is binary (i.e., is scored either 0 or 1), u_g is used for the item score category instead of x_g. The item characteristic function, $P_g(\theta)$, is defined by

$$P_g(\theta) = \text{prob.}[u_g=1|\theta] \ , \qquad [8]$$

and the other operating characteristic for $u_g = 0$, which is denoted by $Q_g(\theta)$, can be written as

$$Q_g(\theta) = \text{prob.}[u_g=0|\theta] = 1 - P_g(\theta) \ . \qquad [9]$$

From Equations 5, 8, and 9

$$I_g(\theta) = [\frac{\partial}{\partial\theta} P_g(\theta)]^2\, [P_g(\theta)\, Q_g(\theta)]^{-1} \ , \qquad [10]$$

which is identical with the item information function defined by Birnbaum (1968), can be obtained for the item information function.

The Constant Information Model

The item characteristic function of the new model, the Constant Information Model (CIM), is given by

$$P_g(\theta) \begin{cases} = \sin^2[a_g(\theta - b_g) + (\pi/4)], \text{ for } \underline{\theta} < \theta < \bar{\theta} \\ = 0 \text{ otherwise}, \end{cases} \qquad [11]$$

where

$$\begin{cases} \underline{\theta} = [-\pi a_g^{-1}/4] + b_g \\ \bar{\theta} = [\pi a_g^{-1}/4] + b_g \end{cases} . \qquad [12]$$

Figure 1
Item Characteristic Functions of Five Binary Items
Following the Constant Information Model

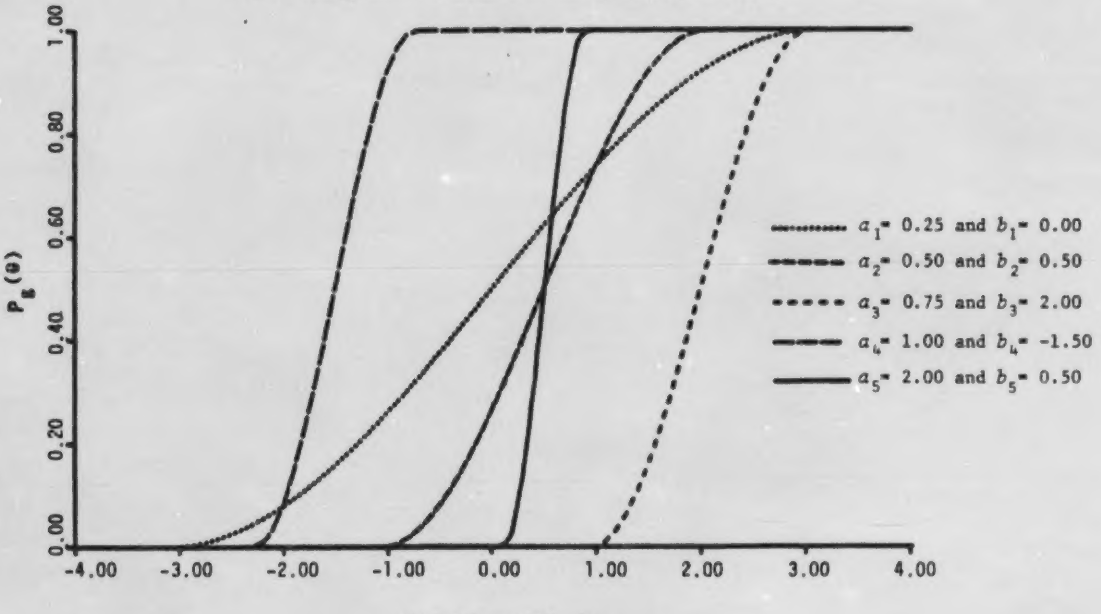


Figure 1 presents five examples of the item characteristic functions in the CIM with varieties of sets of parameters. From Equations 9 and 11, $Q_g(\theta)$ can be written for the other operating characteristic such that

$$Q_g(\theta) \begin{cases} = \cos^2[a_g(\theta - b_g) + (\pi/4)], \text{ for } \underline{\theta} > \theta > \overline{\theta} \\ = 1 \text{ otherwise.} \end{cases} \qquad [13]$$

Since

$$\begin{aligned} \frac{\partial}{\partial\theta} P_g(\theta) &= 2 \sin [a_g(\theta - b_g) + (\pi/4)] \times \\ &\qquad \cos [a_g(\theta - b_g) + (\pi/4)] \times a_g \\ &= 2\, a_g\, [P_g(\theta) Q_g(\theta)]^{\frac{1}{2}} \end{aligned} \qquad [14]$$

can be written for the interval of θ such that

$$[-\pi a_g^{-1}/4] + b_g < \theta < [\pi a_g^{-1}/4] + b_g \;, \qquad [15]$$

it is obvious that in this model:

(A) the item characteristic function is strictly increasing in θ in the above interval of θ,

and

(B)
$$\begin{cases} \lim\limits_{\theta \to \underline{\theta}} P_g(\theta) = 0 \\ \lim\limits_{\theta \to \overline{\theta}} P_g(\theta) = 1 \;. \end{cases} \qquad [16]$$

For convenience, hereafter, the CIM will be considered only for the range of θ given by Equation 15, unless otherwise stated. The mathematical models which satisfy (A) and (B) will be called models of Type I.

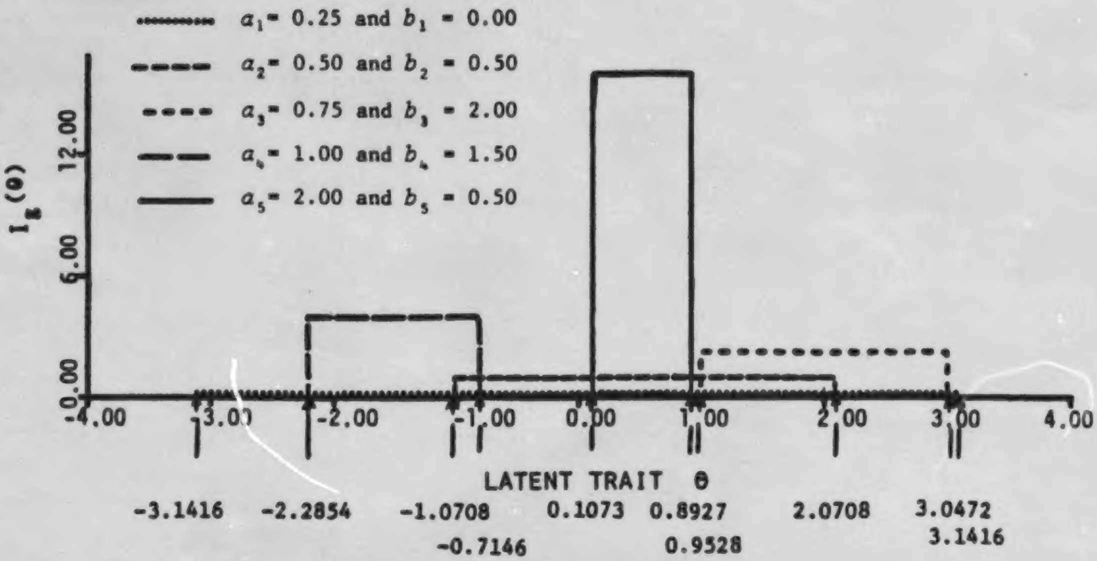
It is obvious from Equations 11, 13, and 14 that the model provides constant item information such that

$$I_g(\theta) = 4a_g^2 = C_g \qquad [17]$$

where C_g indicates the constant amount of information. Figure 2 presents the item information functions for the five items whose item characteristic functions are given in Figure 1. The length of the interval of θ for which the item information function equals C_g is given by

$$\overline{\theta} - \underline{\theta} = \pi C_g^{-\frac{1}{2}} = \Pi[2a_g]^{-1} \qquad [18]$$

Figure 2
Item Information Functions of Five Binary Items
Following the Constant Information Model



The basic function, $A_{x_g}(\theta)$ (Samejima, 1969, 1972), which is defined by

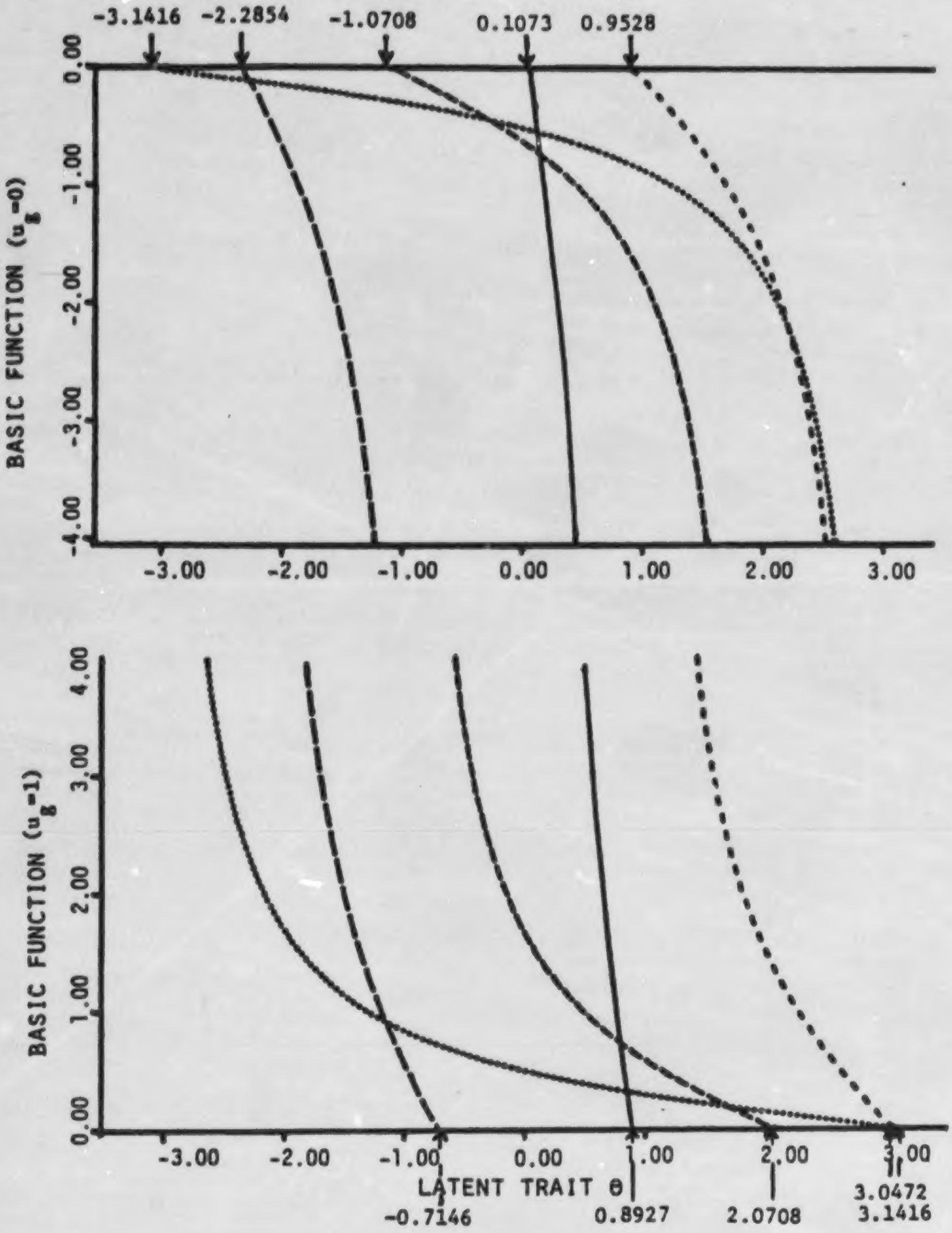
$$A_{x_g}(\theta) = \frac{\partial}{\partial\theta} \log P_{x_g}(\theta) \qquad [19]$$

for the item response category x_g, is obtained for the CIM from Equations 11, 13, 14, and 19 such that

$$A_{u_g}(\theta) \begin{cases} = -2a_g[P_g(\theta)]^{-\frac{1}{2}}[Q_g(\theta)]^{-\frac{1}{2}} \\ = -2a_g \tan[a_g(\theta - b_g) + \pi/4] \\ \quad \text{for } u_g = 0 \\ = 2a_g[Q_g(\theta)]^{\frac{1}{2}}[P_g(\theta)]^{-\frac{1}{2}} \\ = 2a_g \cot[a_g(\theta - b_g) + \pi/4] \\ \quad \text{for } u_g = 1 \end{cases} \qquad [20]$$

Figure 3 presents these basic functions for the five items shown earlier. It is clear from Equation 20 that the basic function for $u_g = 1$ is a strictly decreasing function of θ with positive infinity and zero as its two asymptotes and that for $u_g = 0$ it is a strictly decreasing function of θ with zero and negative in-

Figure 3
Basic Functions of Five Items Following the Constant Information Model for $\underline{u}_g = 0$ (Upper Graph) and for $\underline{u}_g = 1$ (Lower Graph)



finity as its two asymptotes, respectively. This is also confirmed visually by Figure 3.

The item response information functions for each of the binary scores, $u_g = 0$ and $u_g = 1$, are given from Equations 4, 11, 13, and 14 by

$$I_{u_g}(\theta) \begin{cases} = 2a_g^2 \ \sec^2[a_g(\theta - b_g) + \pi/4] \\ = 2a_g^2[Q_g(\theta)]^{-1} > 0 \\ \text{for } u_g = 1 \\ = 2a_g^2 \ \csc^2[a_g(\theta - b_g) + \pi/4] \\ = 2a_g^2[P_g(\theta)]^{-1} > 0 \\ \text{for } u_g = 1 \end{cases} \qquad [21]$$

Figure 4 illustrates these two functions for the item with parameters $a_g = 0.25$ and $b_g = 0.00$, together with the constant item information (=0.25). The response pattern information function, $I_V(\theta)$, is written from Equations 6 and 21 such that

$$I_V(\theta) = 2 \sum_{u_g \varepsilon V} a_g^2[P_g(\theta)]^{-u_g}[Q_g(\theta)]^{u_{g-1}} \qquad [22]$$

and the test information function, $I(\theta)$, is given from Equations 7 and 17 by

$$I(\theta) = 4 \sum_{g=1}^{n} a_g^2 \ . \qquad [23]$$

Figure 5 presents both the set of four response pattern information functions and the test information function for a hypothetical test of two binary items, whose item parameters are $a_1 = 0.25$, $b_1 = 0.00$, $a_2 = 0.50$, and $b_2 = 1.00$, each of which follows the CIM.

It should be noted that the interval of θ for which the item information is a positive constant is always finite. This is related to the fact that

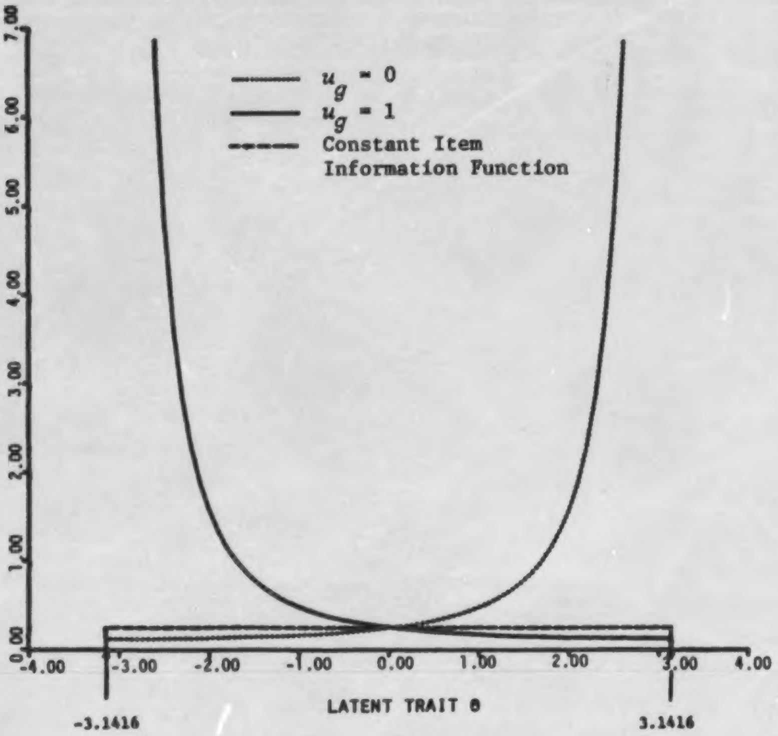
$$\int_{\underline{\theta}}^{\overline{\theta}} [I_g(\theta)]^{\frac{1}{2}} \, d\theta = \pi \ , \qquad [24]$$

for any model of Type I , which includes the CIM (Samejima, 1979). Thus,

$$C_g^{1/2}(\bar{\theta} - \underline{\theta}) = \pi \qquad [25]$$

and $(\bar{\theta} - \underline{\theta})$ must therefore be a finite value.

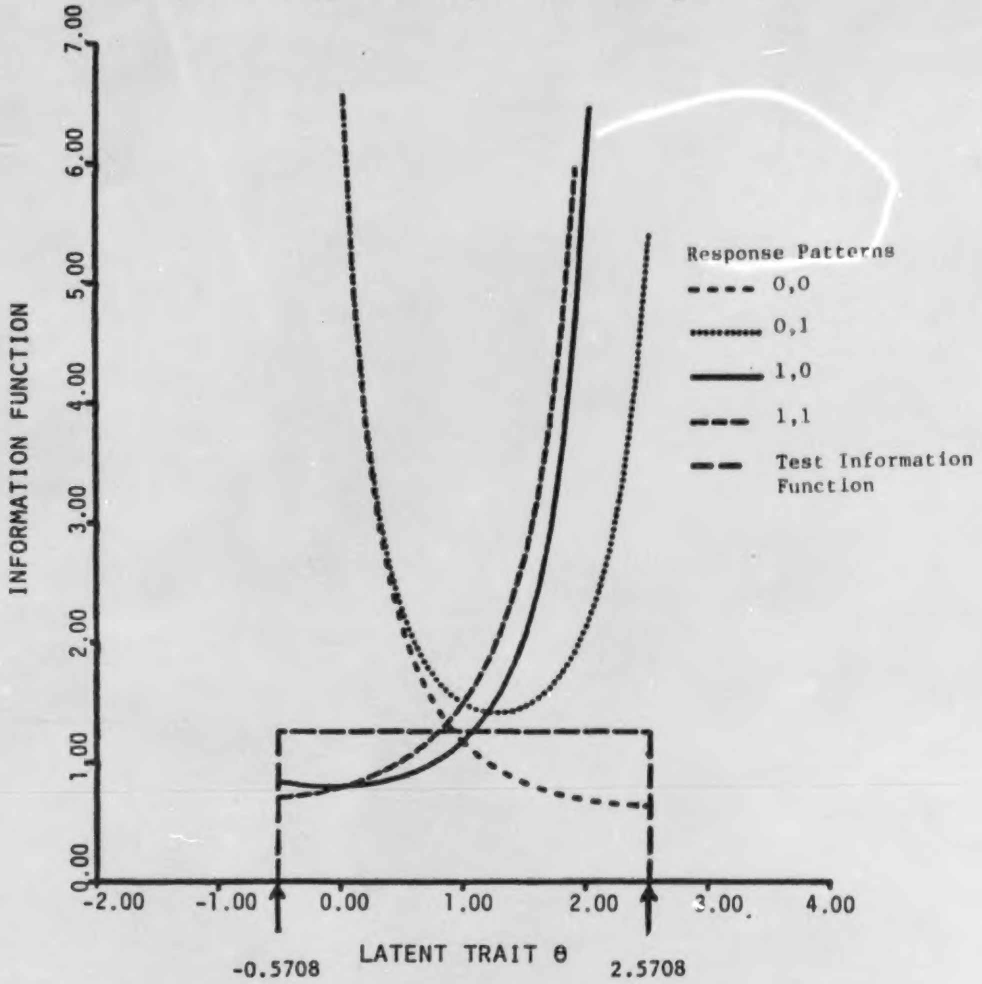
Figure 4
Item Response Information Functions of an Item Following
the Constant Information Model, with the Parameters
$\underline{a}_g = .25$ and $\underline{b}_g = 0.0$ for $\underline{u}_g = 0$ and for $\underline{u}_g = 1$,
Together with the Constant Item Information Function



Use of the Constant Information Model in the Estimation of the Operating Characteristics of Item Response Categories

The methods and approaches for estimating the operating characteristics of the item response categories of a test item developed so far (Samejima, 1977b, 1977d, 1978a, 1978b, 1978c, 1978d, 1978e, 1978f) have common characteristics such that (1) they are made without assuming any prior mathematical form and (2) they use a relatively small number of examinees in the process of estimation. One common restriction in these methods and approaches is that what is needed is the Old Test, consisting of the items whose operating characteristics are known, which provides a constant test information function for the interval of latent trait, or ability, θ, of interest. With this setting, each examinee's ability level is estimated from his/her response pattern by the maximum likelihood estimation; and the set of these maximum likelihood estimates is the main information source of the estimation procedures.

Figure 5
Response Pattern Information Functions of the Four Possible Response Patterns of Two Binary Items Following the Constant Information Model with the Parameters $\underline{a}_1 = .25$, $\underline{b}_1 = 0.0$, $\underline{a}_2 = .50$, and $b_2 = 1.0$



It will be noted that these methods and approaches are useful in a situation when there already is an item pool and it is desired to add more test items to it. In a different setting where the starting point is the very beginning of item pool development, however, these methods appear to be useless, since there is no Old Test, and, consequently, the maximum likelihood estimate of each examinee's ability is not a priori given.

In the above setting, it is very common that a large number of test items, which include many items of intermediate difficulties, is administered to the group of examinees chosen for the purpose of developing the item pool. If, among the items administered, a substantially large subset of binary items can be found which can be regarded as equivalent items, then that subset of items can be used in place of the Old Test. Even though their item characteristic functions are not known, this can be done with the aid of the CIM. In other words, it is possible to expand the estimation methods developed so far to make them usable when there is no Old Test.

Estimation Procedures

In practice, it is necessary to identify these equivalent items without knowing their operating characteristics. This can be done as follows: First of all, the proportion correct must be obtained for each item. If a subset of items exists whose proportions correct are around .5 and close enough to one another, then these items make a good candidate for the subset of equivalent items. Second, a 2 × 2 contingency table, as exemplified in Table 1, must be made for each pair of items in the subset. In order for the items to be equivalent, it is necessary that within the range of sampling fluctuations, these 2 × 2 contingency tables of the bivariate frequency distributions should be symmetric and identical for all the pairs of binary items. This can be checked for every pair of binary items that have passed the first screening, and, very likely, more items must be eliminated through this second screening. Now there can be progression to the third screening of 2^3 contingency tables, to the fourth screening of 2^4 contingency tables, and so forth. Note, however, that an increasingly large number of more complicated contingency tables is usually encountered as progression is made to higher stages of screening. Some criterion must be set for terminating this process, therefore, the remaining items must be accepted by assuming their equivalence. Table 1 illustrates two typical 2 × 2 contingency tables—one is for a pair of equivalent binary items that has a common low discrimination parameter and the other is for a pair that has a common high discrimination parameter.

Table 1
Two Typical 2 × 2 Contingency Tables
for a Pair of Equivalent Items with a
Common Low Discrimination Parameter
and with a Common High Discrimination Parameter

	Item h		
Item g	$u_h=0$	$u_h=1$	Total
Low Discrimination Parameter			
$u_g=0$	110	243	353
$u_g=1$	248	399	647
Total	358	642	1000
High Discrimination Parameter			
$u_g=0$	300	53	353
$u_g=1$	58	589	647
Total	358	642	1000

After the subset of equivalent binary items has been identified, then the

CIM is assumed for each equivalent item. This assumption is made without loss of generality, since the scale of the latent trait is subject to any strictly increasing transformation (Samejima, 1969, 1979).

The next step is to obtain the maximum likelihood estimate, $\hat{\theta}$, of θ on the response pattern of each of the N examinees, with respect to the above subset of equivalent items. Let V* be the examinee's response pattern on the subset of k equivalent items. Since they are equivalent items, the simple test score, t, such that

$$t = \sum_{u_g \in V^*} u_g \qquad [26]$$

is a minimal sufficient statistic, regardless of the model that these item characteristic functions follow (cf. Birnbaum, 1968). Thus, the procedure of maximum likelihood estimation becomes much simpler, using the test score t instead of the response pattern V*. In general, for any model on the dichotomous response level

$$P_{V*}(\theta) = \prod_{u_g \in V*} [P_g(\theta)]^{u_g}[Q_g(\theta)]^{1-u_g} \ . \qquad [27]$$

Since this operating characteristic of the response pattern V* is itself the likelihood function in estimating the examinee's ablity, the symbol $L_{V*}(\theta)$ will be used for this function in the present section. When all the items are equivalent, Equation 27 can be rewritten in the form

$$L_{V*}(\theta) = [P_g(\theta)]^t[Q_g(\theta)]^{k-t} \ . \qquad [28]$$

Thus, for the likelihood equation

$$\frac{\partial}{\partial\theta} \log L_{V*}(\theta) = [\ \frac{\partial}{\partial\theta} P_g(\theta)][t - kP_g(\theta)] \qquad [29]$$

$$[P_g(\theta)Q_g(\theta)]^{-1} = 0 \ ,$$

and the equation

$$t = kP_g(\theta) \qquad [30]$$

is obtained. For the maximum likelihood estimate $\hat{\theta}$,

$$\hat{\theta} = P_g^{-1} \ (t/k) \qquad [31]$$

Now, Equation 11 must be substituted into Equation 31, which results in

$$\hat{\theta} = a_g^{-1} \ [\sin^{-1}(t/k)^{\frac{1}{2}} - \pi/4] + b_g \ . \qquad [32]$$

It is obvious from Equations 15 and 32 that the range of $\hat{\theta}$ is given by

$$[-\pi a_g^{-1}/4] + b_g \le \theta \le [\pi a_g^{-1}/4] + b_g \quad . \qquad [33]$$

Thus, the maximum likelihood estimate of each examinee is obtained, based upon his/her test score for the subset of k equivalent items. This set of N maximum likelihood estimates can be used in place of the one obtained on the Old Test, and the process of the operating characteristic estimation in any combination of a method and an approach can be followed. The error variance, σ^2, is given by

$$\sigma^2 = [kC]^{-1} = [4ka^2]^{-1} \quad , \qquad [34]$$

where

$$C_1 = C_2 = \ldots = C_k = C \qquad [35]$$

and

$$a_1 = a_2 = \ldots = a_k = a \quad . \qquad [36]$$

After the process of estimation has been completed, the latent trait θ can be transformed to another latent trait τ by any strictly increasing function, $\tau(\theta)$. To give some examples, if θ is transformed to τ by

$$\tau = P_g^{*-1}([\sin\{a_g(\theta-b_g) + (\pi/4)\}]^2) \quad , \qquad [37]$$

where

$$P_g^*(\tau) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{a_g^*(\tau - b_g^*)} \exp[-t^2/2]dt \quad , \qquad [38]$$

then all the equivalent items will follow the normal ogive model specified by Equation 38, and

$$\begin{cases} \underline{\theta} = -\infty \\ \overline{\theta} = \infty \; ; \end{cases} \qquad [39]$$

if the transformation is such that

$$\tau = (\beta_g-\alpha_g) \sin^2[a_g(\theta - b_g) + (\pi/4)] + \alpha_g \quad , \qquad [40]$$

then they will follow the linear model, whose item characteristic is given by

$$P_g(\theta) = (\theta-\alpha_g)(\beta_g-\alpha_g)^{-1} \qquad [41]$$

with

$$\begin{cases} \underline{\theta} = \alpha_g \\ \overline{\theta} = \beta_g \end{cases} \qquad [42]$$

and if θ is transformed to τ by

$$t = [1/(Da^*_g)] \log[\tan^2\{a_g(\theta - b_g) + (\pi/4)\}] + b^*_g \quad , \qquad [43]$$

then they will follow the logistic model, which is characterized by

$$P^*_g(\tau) = [1 + \exp\{-Da^*_g(\tau - b^*_g)\}]^{-1} \qquad [44]$$

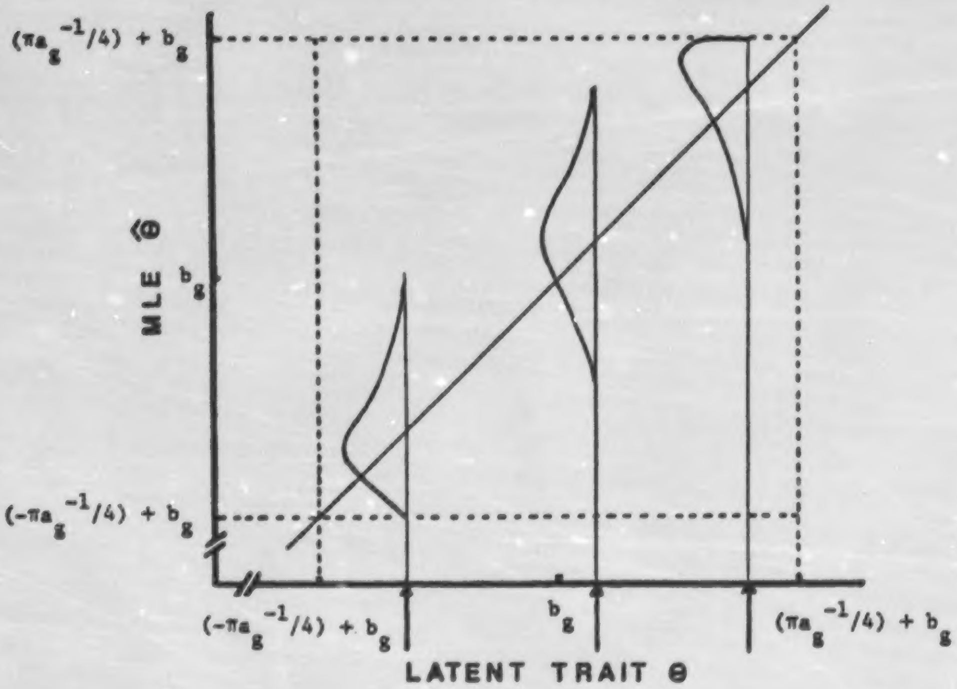
with the $\underline{\theta}$ and $\overline{\theta}$ given in Equation 39. Similar transformations can be made to change the item characteristic functions of $\underline{k}$ equivalent items from those of the CIM to those of any other models of Type I. In each case, the newly estimated operating characteristics of the other items will be transformed according to the specific transformation of the latent trait θ to τ.

Some Necessary Considerations

In using the generalized method of the operating characteristic estimation, which was described in the preceding section, a few problems must be considered. First of all, the constant test information provided by the subset of equivalent binary items with the CIM should be substantially large, so that the normal approximation for the conditional distribution of $\hat{\theta}$, given θ, will be acceptable. On the other hand, a substantially wide range of ability θ, for which the test information is constant, is needed in order to make the estimation of the operating characteristics of the other items in the item pool meaningful. These two are opposing factors, as is obvious from Equations 15 and 17. The solution for this problem is to use a substantially large number of equivalent binary items, whose common discrimination parameter is low.

Another problem is the effect of the range of $\hat{\theta}$ on the speed of convergence of the conditional distribution of $\hat{\theta}$, given θ, to the normal distribution, $\underline{N}(\theta, \{kC\}^{-\frac{1}{2}})$. Since the range of $\hat{\theta}$ is a finite interval which is given by Equation 33, it should be expected that the truncation of the conditional distribution makes the convergence slow around the values of θ close to $(-\pi \underline{a}_g^{-1}/4) + b_g$ and $(\pi \underline{a}_g^{-1}/4) + \underline{b}_g$, as is illustrated in Figure 6. A solution for this problem is again to use a subset of equivalent binary items whose common discrimination parameter is low so that the range of θ is wide enough to include all the examinees sufficiently inside of the two endpoints of the interval of θ. An alternative for the above solution is to exclude examinees whose $\hat{\theta}$'s are close to $(-\pi \underline{a}_g^{-1}/4) + \underline{b}_g$ or $(\pi \underline{a}_g^{-1}/4) + \underline{b}_g$. In the second solution, however, the number of examinees usable for the estimation will be decreased, and this may substantially affect the accuracy of the estimation of the operating characteristics. It is worth noting that the solution for the first problem also seems to be the best solution for the second problem.

Figure 6
Graphical Illustration of the Conditional Density
Functions of the Maximum Likelihood Estimate $\hat{\theta}$,
Given the Latent Trait θ

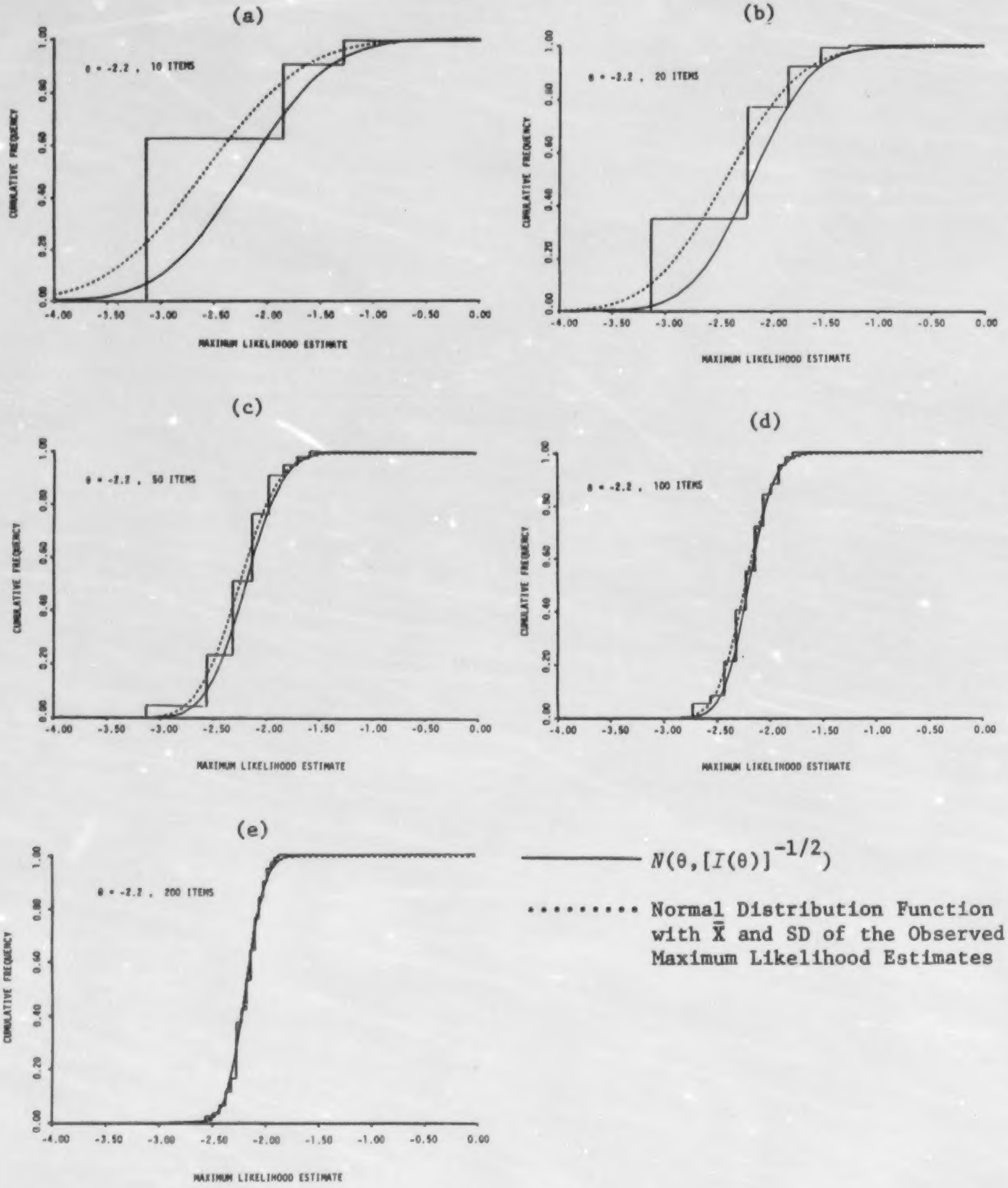


Monte Carlo Study

Method. To pursue the speed of convergence that the conditional distribution of the maximum likelihood estimate, $\hat{\theta}$, approaches normality, $N(\theta, I(\theta)^{-\frac{1}{2}})$, a monte carlo study was done. One hundred hypothetical examinees were assumed to exist at each of eight different levels of ability θ, i.e., -3.0, -2.2, -1.4, -0.6, 0.2, 1.0, 1.8, and 2.6. It was assumed that each examinee had taken 20 sessions of equivalent tests, in each of which 10 equivalent binary items were given. Each binary item was assumed to follow the CIM with the parameters a_g = .25 and b_g = 0.00, and the response pattern was calibrated by the monte carlo method for each examinee in each of the 20 hypothetical sessions of testing. The cumulative test score was calculated after each session by summing all the binary item scores that the examinee had obtained. In this way, the number of the binary items used for the computation of the cumulative test score was 10 after the first session, 20 after the second session, 30 after the third session, and 200 after the completion of the twentieth session. The maximum likelihood estimate of ability was obtained for each of the 800 examinees, based upon each cumulative test score following the method described in the preceding section.

Results. Figure 7 presents the resultant cumulative frequency distribution

Figure 7
Cumulative Frequency Distributions of the 100 Maximum Likelihood Estimates for θ=-2.2 for Group 2, Based on 10, 20, 50, 100, and 200 Items



of the maximum likelihood estimates of the 100 examinees whose ability level was -2.2 (i.e., relatively close to the lower endpoint, $-\pi$, of the interval for which the item information, $I_g(\theta)$ assumes the positive value, $-.25$), after the completion of each of sessions 1, 2, 5, 10, and 20. In each graph, the normal distribution functions are also presented with θ ($= -2.2$) and $I(\theta)^{-\frac{1}{2}}$ as the two parameters and with the mean and the standard deviation of the observed 100 maximum likelihood estimates. As was expected, the cumulative frequency distribution shows a substantial skewness to the positive side, when the number of binary items is as small as 10; and therefore the normal distribution function with the two empirically obtained parameters provides a curve that is located further to the right side of $N(\theta, I(\theta)^{-\frac{1}{2}})$. This tendency still holds, though slightly, even when the number of binary items is as large as 100.

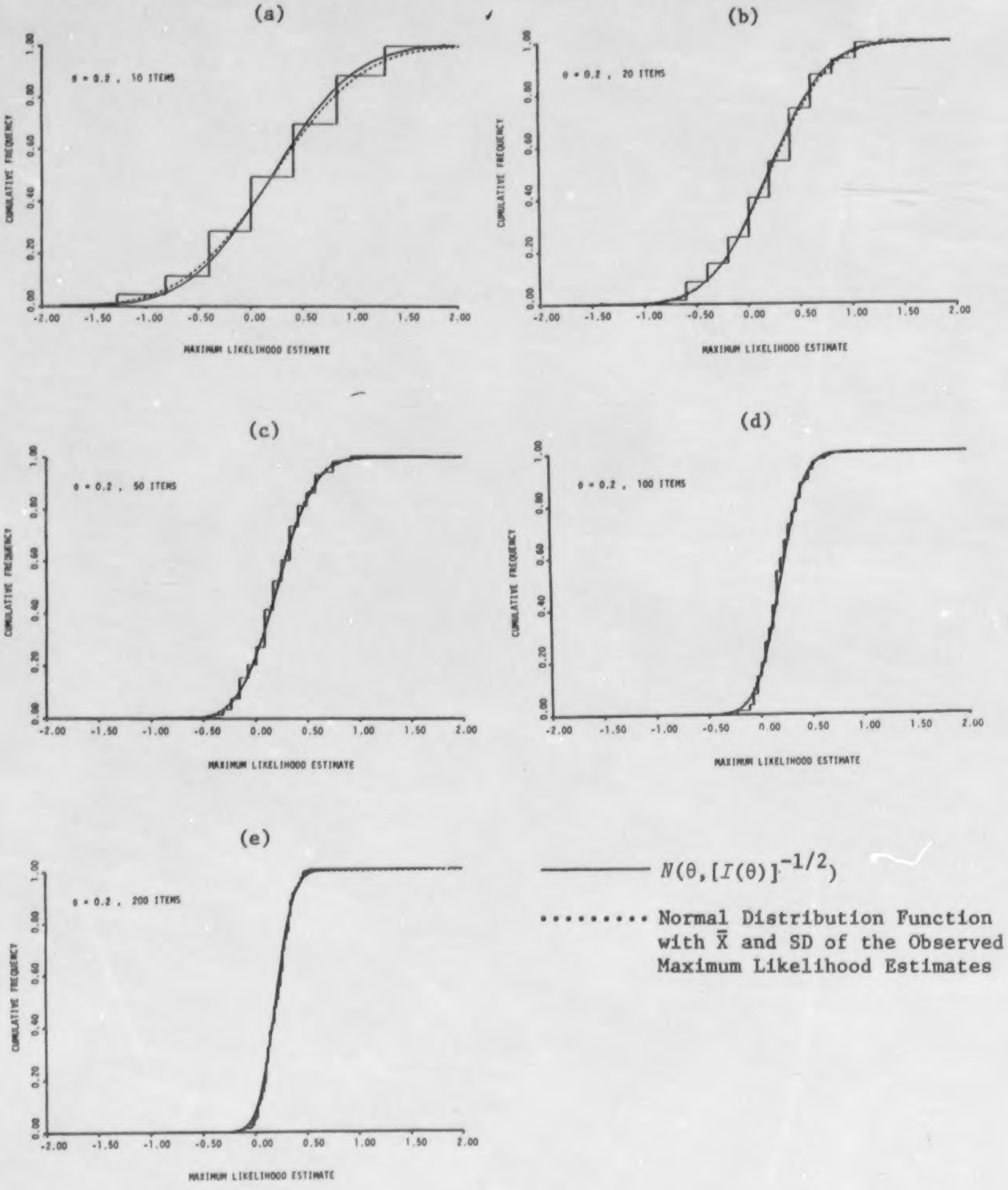
For the purpose of comparison, Figure 8 presents a similar set of graphs for another group of 100 hypothetical examinees whose ability levels were uniformly .2 (i.e., a value which is far from the two endpoints, $-\pi$ and π, of the interval for which the item information function of each equivalent binary item assumes the positive constant). The results illustrated in Figure 8 make a good contrast with those in Figure 7, in which the two normal distribution curves almost overlap each other, even when the number of items used for obtaining the maximum likelihood estimate is as small as 20. This indicates a much faster convergence of the conditional distribution of the maximum likelihood estimate to normality with θ and $I(\theta)^{\frac{1}{2}}$ as the two parameters, in comparison with the case in which the ability level is closer to one of the two endpoints of the interval, $(-\pi, \pi)$.

If the latent trait θ is transformed to τ through Equations 38 and 39 so that each equivalent binary item follows the normal ogive model with $a_g^* = 1.00$ and $b_g^* = 0.00$, then the values of τ corresponding to $\theta = -2.2$ and $\theta = 0.2$ are, approximately, -1.60 and $.13$. If Equation 44 is used for the transformation so that each equivalent binary item follows the logistic model with the same parameters—a_g^* and b_g^*—with the scaling factor $D = 1.7$, then these corresponding values of τ are approximately -1.68 and $.12$, respectively. A similar transformation by means of Equation 41, which provides the linear model with the parameters $\alpha_g = -2.5$ and $\beta_g = 2.5$ for each of the equivalent binary items, results in -2.23 and $.25$ as the approximate values of τ corresponding to $\theta = -2.2$ and $\theta = 0.2$, respectively.

Discussion and Conclusion

A new model for a binary item, the Constant Information Model, has been introduced, and its characteristics and usefulness have been discussed. As was pointed out (Samejima, 1975, 1977a, 1977b, 1977c, 1977d, 1978a, 1978b, 1978c, 1978d, 1978e, 1978f), latent trait theory enlarges its horizon if full use is made of information functions, enabling types of research to be conducted which could not otherwise be done. For this reason, the Constant Information Model will contribute to the productivity of research in the area of computerized adaptive testing as well as in other areas, as exemplified in this paper.

Figure 8
Cumulative Frequency Distributions of the 100 Maximum Likelihood Estimates for θ=.2 for Group 5, Based on 10, 20, 50, 100, and 200 Items



REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, 34 (4, Pt. 2, Monograph No. 17)

Samejima, F. A general model for free-response data. Psychometrika Monograph Supplement, 1972, 37 (1, Pt. 2, Monograph No. 18)

Samejima, F. Graded response model of the latent trait theory and tailored testing. In C. L. Clark (Ed.), Proceedings of the first conference on computerized adaptive testing. (U.S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

Samejima, F. A use of the information function in tailored testing. Applied Psychological Measurement, 1977, 1, 233-247. (a)

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 1977, 42, 163-191. (b)

Samejima, F. The application of graded response models: The promise of the future. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1978. (c)

Samejima, F. Estimation of the operating characteristics of item response categories I: Introduction to the two-parameter beta method (Research Report 77-1). Knoxville: University of Tennessee, Department of Psychology, 1977. (d)

Samejima, F. Estimation of the operating characteristics of item response categories II: Further development of the two-parameter beta method (Research Report 78-1). Knoxville: University of Tennessee, Department of Psychology, 1978. (a)

Samejima, F. Estimation of the operating characteristics of item response categories III: The normal approach method and the Pearson system method (Research Report 78-2). Knoxville: University of Tennessee, Department of Psychology, 1978. (b)

Samejima, F. Estimation of the operating characteristics of item response categories IV: Comparison of the different methods (Research Report 78-3). Knoxville: University of Tennessee, Department of Psychology, 1978. (c)

Samejima, F. Estimation of the operating characteristics of item response categories V: Weighted sum procedure in the conditional P.D.F. approach (Research Report 78-4). Knoxville: University of Tennessee, Department of Psychology, 1978. (d)

Samejima, F. Estimation of the operating characteristics of item response categories VI: Proportioned sum procedure in the conditional P.D.F. approach (Research Report 78-5). Knoxville: University of Tennessee, Department of Psychology, 1978. (e)

Samejima, F. Estimation of the operating chracteristics of item response categories VII: Bivariate P.D.F. approach with normal approach method (Research Report 78-6). Knoxville: University of Tennessee, Department of Psychology, 1978. (f)

Samejima, F. Constant information model: A new, promising item characteristic function (Research Report 79-1). Knoxville: University of Tennessee, Department of Psychology, 1979.

DISCUSSION: SESSION 4

ROBERT TSUTAKAWA
UNIVERSITY OF MISSOURI

In reading Samejima's paper, one soon realizes that her ideas are quite different and provocative: She appears to be knocking on the door of the foundations of latent trait theory. In my discussion of her paper I will attempt to provide appropriate motivation for the Constant Information Model (CIM) and point out its relation to other statistical methods, reviewing some of the important issues and raising points that could be discussed further.

The problem that motivates the CIM is not unique to mental testing. Generally, in estimating a parameter θ, the estimator will have a variance depending on the unknown parameter θ. When estimating several parameters $\theta_1,\ldots,\theta_N$, say, the ability of N people, there will be estimators with different variances. (An important exception to the general property is the normal linear model where the constancy of variance plays an important role.) If interest is in making statistical inferences based on the estimated values, the constancy of variance will open up a variety of statistical procedures, such as the analysis of variance. Moreover, in designing an experiment the constancy of variance will permit the running of an experiment with guaranteed precision.

Consider the item characteristic function of the CIM given by

$$P_g(\theta) = \sin^2 [a_g(\theta - b_g) + \pi/4], \quad \underline{\theta}_g < \theta < \overline{\theta}_g \ . \qquad [1]$$

If its value is denoted by p, the inverse transformation is

$$\theta = a_g^{-1} \sin^{-1} \sqrt{p} - \frac{\pi}{4a_g} + b_g, \quad 0 < p < 1 \ . \qquad [2]$$

In this form it can be noted that this is essentially the arc sine transformation, which was used in the 1930s and 1940s to stabilize the variance of binomial proportions and to obtain a better normal approximation. The arc sine transformation is also used by Bayesians for binomial samples to achieve likelihood functions that are "data translated" and was used by Jeffereys to obtain the noninformative prior for an unknown proportion. Samejima's transformation should thus be of particular interest to Bayesians.

Regarding the range of θ in the CIM, it is disconcerting to restrict θ to intervals depending on the item. It seems more realistic to extend the range to the whole real line by defining it as

$$P_g(\theta) = \begin{cases} 0 \text{ if } \theta \leq \underline{\theta}_g \\ \text{as above if } \underline{\theta}_g < \theta < \overline{\theta}_g \\ 1 \text{ if } \theta \geq \overline{\theta}_g \text{ .} \end{cases} \qquad [3]$$

With this extension, different response functions can be dealt with, allowing for the probability of a correct response on a given item, g, to be 1 whenever the ability θ exceeds $\overline{\theta}_g$ and 0 whenever it is less than $\underline{\theta}_g$. Note that when $\theta \geq \overline{\theta}_g$ or $\theta \leq \underline{\theta}_g$, the formal information is 0. However, the experimenter may have some idea about whether θ is very low or very high.

Given a fixed number of items n, equivalent tests are not only a convenience but a practical necessity in order to attain constant total information over all θ in the range of interest. With nonequivalent tests the total information will generally (with rare exceptions) depend on θ. This raises the important question about how to find a set of equivalent items.

There are two simple ways of constructing tests so the items are equivalent. One is to use the n items in random order. In this case the probability of a correct response is constant for any ability θ; the assumption of local independence, however, is violated. Another method is to select the n items at random from a large pool. In this case, the probability of a correct response for ability θ is equal to the average probability of correct response (averaged over the whole pool), and the assumption of local independence can be defended. If the items are very different, however, Bayesians would consider this poor practice.

Given a set of n nonrandom items, testing their equivalence is a challenging statistical problem. Although Samejima has given some suggestions, considerably more work is needed before these suggestions can be put into practice.

There are implications for the use of the Constant Information Model in tailored testing. It seems reasonable to start with items with low a_g at the beginning when the location of θ is uncertain, and then use those with higher a_g as the region in which θ is likely to belong is narrowed. With a good selection rule, this is likely to be more efficient than having a large number of equivalent items with low a_g. The gain in efficiency will more than offset the inconvenience of not having constant total information.

SESSION 5:

ITEM LINKING AND EQUATING

A TEST OF THE ADEQUACY OF CURVILINEAR SCORE EQUATING MODELS

GARY MARCO, NANCY PETERSEN,
AND ELIZABETH STEWART
EDUCATIONAL TESTING SERVICE

THE EFFECTS OF CONTEXT ON LATENT TRAIT MODEL ITEM PARAMETER AND TRAIT ESTIMATES

WENDY M. YEN
CTB/MCGRAW HILL

EFFECTS OF SAMPLE SIZE ON LINEAR EQUATING OF ITEM CHARACTERISTIC CURVE PARAMETERS

MALCOLM J. REE AND
HARALD E. JENSEN
AIR FORCE HUMAN RESOURCES
LABORATORY

DISCUSSION

GAIL IRONSON
BOWLING GREEN STATE
UNIVERSITY

A TEST OF THE ADEQUACY OF CURVILINEAR SCORE EQUATING MODELS

GARY MARCO, NANCY PETERSEN, AND ELIZABETH STEWART
EDUCATIONAL TESTING SERVICE

In many common testing situations it is necessary to compare the test scores of examinees who have taken different forms of a test. In practice, two forms of a test cannot be expected to be of exactly equal difficulty for examinees at all ability levels. Therefore, a comparison of raw scores on two forms of a test will be unfair to the examinees who have taken the more difficult form. Statistical procedures that have been developed to deal with this problem are referred to as equating methods.

In an ideal psychometric world, tests on which scores need to be equated would be parallel in all important respects: An anchor test, if used, would be parallel to the total tests, and random samples on which to base the equating would always be available. In actual testing practice, however, scores must sometimes be equated under less than optimum conditions. This study is the first part of a larger study, the purpose of which is to examine the adequacy of score-equating models when certain sample and test characteristics are systematically varied. The emphasis in this part of the study is on curvilinear models, whereas the second part focuses on linear models. This study is more comprehensive than previous studies of equating models (e.g., Levine, 1955; Rentz & Bashaw, 1975; Slinde & Linn, 1977, 1978; Tucker, 1974) in that it includes a greater variety of equipercentile, linear, and ICC models and investigates equatings based on dissimilar samples as well as on random samples.

EQUATING MODELS

An equating method is an empirical procedure for determining a transformation to be applied to the scores on one of two forms of a test. Its purpose is, ideally, to transform the scores in such a way that it makes no difference to the examinee which form of the test he or she takes. This ideal can be reached only if (1) the two forms of the test measure exactly the same latent trait (ability or skill) and yield scores that are equally reliable and (2) the equating transformation is invertible.

Because an equating method is an empirical procedure, it involves a design for data collection and a rule for determining the transformation. There are three basic designs for data collection and three general rules for determining the transformation. Any of the three designs for data collection can be used with any of the three transformation rules.

Data Collection Designs

The three designs for data collection are the single-group method, the equivalent group method, and the anchor test method (Lord, 1975). All of the equating procedures used in this study assume that the data were collected using the anchor test method. An anchor test design requires administering one form of a total test to one group of examinees, a second form to a second group of examinees, and a common anchor test to both groups. The anchor test can be either internal or external to the tests to be equated. The anchor test is used to reduce equating bias resulting from differences in ability between the two groups.

Transformation Rules

The three general rules for determining the transformation are:

1. Equipercentile equating. Choose a transformation such that scores from the two tests will be equated if they correspond to the same percentile rank in some group of examinees.
2. Linear equating. Choose a linear transformation such that scores from the two tests will be equated if they correspond to the same number of standard deviations from the mean in some group of examinees.
3. Item characteristic curve (ICC) equating. Choose a transformation such that true scores from the two tests will be equated if they correspond to the same estimated level of the latent trait underlying both tests.

All three types of equating were represented in this study.

Equipercentile and linear methods using anchor test data can be further classified as to whether the equating is done directly or indirectly by frequency estimation. In direct equipercentile equating, scores on each test and on the anchor test are first equated separately within each group. Then, scores on the two tests to be equated are said to be equivalent if they correspond to the same score on the anchor test. Frequency estimation (Angoff, 1971, pp. 581-582), on the other hand, makes use of the combined distribution of scores on the anchor test. The score distributions of the tests to be equated are estimated for the combined group of examinees, and these estimated distributions are then used as if they had been observed from a single-group design. The resulting marginal distributions on the two forms are used in the case of equipercentile equating; the resulting estimated means and standard deviations on the two forms are used in the case of linear equating.

Operationalizing the Models

Many of the linear methods require error variance estimates. The three methods of estimating error variances that were used in this study were Angoff's (1953) method, which uses anchor test data; Feldt's method (1975), which uses part-test data; and coefficient alpha, which uses item-response data.

Table 1
Description of Equating Models

Major Assumptions	Data	Error Variances Known: External Anchor	Error Variances Known: Internal Anchor	Codes: External Anchor	Codes: Internal Anchor	Model	Reference
I. Linear Observed-Score Models							
A. Constancy of Regression							
1. X on $\underline{v}$ and Y on $\underline{v}$							
a. Curvilinear	Bivariate Distributions	None	None	FE-O	FE-O	Frequency Estimation	Angoff, 1979
b. Linear	Cross-Products Matrices	None	None	T1	T1	Tucker 1	Angoff, 1971
2. X on $\underline{v}'$ and Y on $\underline{v}'$ are linear	Cross-Products Matrices	$\underline{v}_a$, $\underline{v}_b$	$\underline{v}_a$, $\underline{v}_b$	T2O-E[a]	T2O-I[b]	Tucker 2 Observed	Tucker, 1974
3. V on $\underline{x}'$ and V on $\underline{y}'$ are linear	Cross-Products Matrices	$\underline{x}_a$, $\underline{y}_b$	$\underline{x}_a$, $\underline{v}_a$, $\underline{y}_b$, $\underline{v}_b$	T3O-E[a]	T3O-I[b]	Tucker 3 Observed	Tucker, 1974
B. $\underline{y}$ and V and Y and V Congoneric and Constancy of Regression of X′ on $\underline{v}'$ and of Y′ and $\underline{v}'$	Cross-Products Matrices	$\underline{x}_a$, $\underline{v}_b$, $\underline{y}_b$, $\underline{v}_b$	$\underline{x}_a$, $\underline{v}_a$, $\underline{y}_b$, $\underline{v}_b$	LER[a]	LER[b]	Levine Equally Reliable	Levine, 1955
C. Distributional Assumptions							
1. X|$\underline{v}$ and Y|$\underline{v}$ are normal	Cross-Products Matrices	None	None	PAC[h]	PAC[i]	Potthoff A	Potthoff, 1966
2. V|$\underline{x}$ and V|$\underline{y}$ are normal	Cross-Products Matrices	None	None	PB	PB[f,g]	Potthoff B	Potthoff, 1966
3. X and V and Y and V are bivariate normal	Cross-Products Matrices	None	None	PAC[h]	PAC[i]	Potthoff C	Potthoff, 1966
4. X and V and Y and V are more general than bivariate normal to allow for group differences	Cross-Products Matrices	None	None	PD	PD	Potthoff D	Potthoff, 1966
II. Curvilinear Observed-Score Models							
A. Constancy of Regression of X on $\underline{v}$ and Y on $\underline{v}$ are curvilinear	Bivariate Distributions	None	None	FE-EQ	FE-EQ	Frequency Estimation	Angoff, 1971
B. No explicit assumptions	Marginal Distributions	None	None	EQ	EQ	Equipercentile	
III. Linear True Score Models							
A. Constancy of Regression							
1. X′ on $\underline{v}'$ and Y′ on $\underline{v}'$							
a. Curvilinear	Bivariate Distributions	$\underline{x}_a$, $\underline{y}_b$	$\underline{x}_a$, $\underline{y}_b$	FE-T	FE-T	Frequency Estimation	Angoff, 1971
b. Linear	Cross-Products Matrices	$\underline{x}_a$, $\underline{v}_a$, $\underline{y}_b$, $\underline{v}_b$	$\underline{x}_a$, $\underline{v}_a$, $\underline{y}_b$, $\underline{v}_b$	T2T-E[c]	T2T-I[d]	Tucker 2 True	Tucker, 1974
2. V′ on $\underline{x}'$ and V′ on $\underline{y}'$ are linear	Cross-Products Matrices	$\underline{x}_a$, $\underline{y}_b$	$\underline{x}_a$, $\underline{v}_a$, $\underline{y}_b$, $\underline{v}_b$	T3T-E[c]	T3T-I[d]	Tucker 3 True	Tucker, 1974

(continued on next page)

Table 1, continued

Major Assumptions	Data	Error Variances Known		Codes		Model	Reference
		External Anchor	Internal Anchor	External Anchor	Internal Anchor		
B. X and V and Y and V Congeneric							
1. Constancy of regression of X' on v' and of Y' on v'	Cross-Products Matrices	x_a, v_a, y_b, v_b	x_a, v_a, y_b, v_b	LUR[e]	LUR[f]	Levine Unequally Reliable	Levine, 1955
2. Observed scores on test form are linear function of observed scores on ideal test. True scores on ideal test are multiplicative function of true scores on anchor test	Cross-Products Matrices	x_a, v_a, y_b, v_b	None	TML-E[e]	TML-E[f,g]	Tucker Modified Levine	Tucker, 1974
3. True scores on test form are linear function of true scores on anchor test							
a. As stated above	Cross-Products Matrices	x_a, y_b	x_a, y_b,	LXY-E[e]	LXY-I[f]	Lord X Y	Lord, 1976
b. As stated above	Cross-Products Matrices	v_a, v_b	v_a, v_b	LV-E[e]	LV-I[f]	Lord V	Lord, 1976
c. As stated above	Cross-Products Matrices	x_a, v_a, y_b, v_b	x_a, v_a, y_b, v_b	LML-E[e]	LML-I[f]	Lord Maximum Likelihood	Lord, 1976 Birch, 1964
4. True scores on part test are linear function of true scores on another test	Cross-Products Matrices	None		LCS		Lord Congeneric Subtests	Lord, 1976
IV. Curvilinear True Score Models							
A. X and V and Y and V Congeneric							
1. Item response function represented by a 1-parameter logistic function	Item Responses	None	None	ICC-1	ICC-1	1-Parameter Logistic ICC	Lord, 1975
2. Item response function represented by a 3-parameter logistic function	Item Responses	None	None	ICC-3	ICC-3	3-Parameter Logistic ICC	Lord, 1975

[a-f] These models yield identical conversion parameters when the Angoff method of estimating error variances is used.

[g-i] These models always yield identical conversion parameters.

Notes. x, y, and v are the observed score on the new form, the old form, and the anchor test, respectively.
x', y', and v' are, respectively, the true scores.
a, b, and c are the group taking tests X and V, the group taking tests Y and V, and the combined group, respectively.

One of the ICC procedures utilized the ICC parameters (item difficulties) from the 1-parameter logistic test model; the other utilized the ICC parameters from the 3-parameter logistic test model. The computer program LOGIST (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976) was used to estimate item parameters and examinee abilities. For both models, true formula scores (R-W/4) were equated by calculating the true formula scores on each test form corresponding to selected ability levels (Lord, 1975) and interpolating as necessary. Since for either model there is a functional relationship between ability and true score, the true formula score is readily computed by the equation

$$R_g - (N - R_g)/(A - 1), \qquad [1]$$

where

R is the true number-correct score at ability g computed by summing the item proportions correct under the model,
N is the number of test items, and
A is the number of response options for the items in the test (five choices in all cases).

The various equating models used in this study are described briefly in Table 1. The models are categorized by whether the procedure results in a linear or curvilinear transformation between observed or true scores. The table provides references and information as to the major assumption underlying each model, the kind of data required, and whether specific error variance estimates are needed. If the codes for the model differ for an external and an internal anchor test, then the formulas for computing the transformation parameters differ in the two cases. A total of 40 linear (including 4 based on the marginal means and standard deviations resulting from frequency estimation), 2 equipercentile, and 2 ICC equating models were used in the study.

STUDY DESIGN

Computer files for two national administrations (April 1975 and November 1975) of the verbal portion of the College Board Scholastic Aptitude Test were obtained from Educational Testing Service (ETS). Fourteen pairs of total test scores were equated on the basis of the data from each of the two administrations. Records of test scores, responses to test items, and responses to a Student Descriptive Questionnaire (SDQ) were accessed to construct samples having specified characteristics, to extract or to calculate selected scores for a number of special purpose total tests and anchor tests, and to compute item statistics and other data needed for the various equating models.

Equating Design

The combinations of total test and anchor test used in the various equatings for each of the two administrations can be classified into five categories, referred to here as test variations. (There were 11 test variations used in the full study of which this study is a part.)

Equating a Test to Itself

In this part of the study (see Table 2), a single form of the operational verbal portion of the SAT (SAT-V) was treated as if it represented two different forms; that is, it was to be equated to itself. (This type of design was first used by Levine, 1955.) This part of the study was designed to investigate the effects of varying (1) the relative difficulty levels of the total test and the anchor test and (2) the degree of similarity between the two samples on which the equating operations were based.

Table 2
Design for Equating a Medium-Difficulty Test
to Itself through an Anchor Test of Similar Content

Test Variation	Anchor Test		Relation Between Samples	SAT-Verbal Score Level	
	Location	Difficulty		New Form Sample	Old Form Sample
1	External	Medium	Random	Middle	Middle
			Dissimilar	Middle	High
4	Internal	Medium	Random	Middle	Middle
			Dissimilar	Middle	High
5	Internal	Easy	Random	Middle	Middle
			Dissimilar	Middle	High
	Internal	Hard	Random	Middle	Middle
			Dissimilar	Middle	High

SAT-V was equated to itself through two anchor tests that, like the total test, were of medium difficulty. One was external (Test Variation 1), and one was internal (Test Variation 4). SAT-V was also equated to itself through two internal anchor tests that differed from it in difficulty (Test Variation 5). One of the internal anchor tests was easier than SAT-V; and the other, more difficult.

For each administration a single pair of random samples and a single pair of dissimilar samples were used for all equatings of SAT-V to itself. The dissimilar samples, by virtue of the sample selection procedure, were expected to be of middle and high verbal ability, respectively.

Equating Tests of Different Difficulty

In this part of the study (see Table 3), three total tests were constructed for each administration from a pool comprising the operational SAT-V items and the items in a nonoperational section of verbal material. The purpose of this part of the study was to examine the effects of varying (1) the relative difficulty levels of the two total tests on which scores were to be equated and (2) the degree of similarity between the two equating samples.

Table 3
Design for Equating Tests that Differed Only in Difficulty through an Internal Anchor Test of Similar Content and Medium Difficulty

Test Variation	Total Test Difficulty		Relation Between Samples	SAT-Verbal Score Level	
	New Form	Old Form		New Form Sample	Old Form Sample
8	Easy	Medium	Random	Middle	Middle
			Dissimilar	Low	Middle
	Medium	Hard	Random	Middle	Middle
			Dissimilar	Middle	High
9	Easy	Hard	Random	Middle	Middle
			Dissimilar	Low	High

Pairs of total tests constructed to differ in difficulty were equated through an internal anchor test of medium difficulty. The random and dissimilar samples used in equating SAT-V to itself were used in these equatings also, along with an additional sample expected to be of low verbal ability.

In Test Variation 8 an easy test was equated to a medium-difficulty test and the medium-difficulty test was equated to a hard test. In Test Variation 9 the easy test was equated to the hard test. For equatings based on dissimilar samples, data from the low-ability sample was used for the easy test; from the middle-ability sample, for the medium-difficulty test; and from the high-ability sample, for the hard test.

Tests

Several scores calculated as part of the normal processing for SAT administrations were included in the study. Included also were scores computed on a number of tests constructed retrospectively solely for use in the study. The general approach followed in constructing the special purpose tests entailed (1) developing content and statistical specifications for all special purpose tests, (2) identifying sets of items in accordance with those specifications, and (3) identifying subsets of items on which separate scores were to be obtained for use in calculating one of the three sets of reliability estimates required for the equating analyses.

Test content was specified only in terms of distributions of item types, although more detailed specifications are followed in developing operational forms of SAT-V. SAT-V is composed of three types of discrete five-choice items--antonyms (25 items), analogies (20 items), and sentence completions (15 items)--and of five reading comprehension passages, each of which is followed by 5 five-choice items based on the passage.

Statistical specifications were stated in terms of the item statistics conventionally used in the development of ETS tests (described by Angoff & Dyer, 1971, pp. 9-10). The equated delta (Δ_e) served as the index of item difficulty; and the biserial correlation ($\underline{r}_b$) of the item score with the total score on the operational test in which the item appeared, as the index of item discrimination. The statistic Δ_e is an estimate of the difficulty of the item for a standard reference group. It ranges in value from about 6 (very easy) to 18 (very hard). If a test composed of five-choice items were of middle difficulty for the reference group, its mean Δ_e would be 12.0.

The item statistics used to construct the tests were taken from the results of item analyses routinely conducted for each new form of the SAT-V. The analyses were based on systematic samples of approximately 1,700 to 2,000 examinees each. After all full-length total tests and anchor tests had been identified, part tests for use in reliability estimation (Feldt, 1975) were created. Each part test was parallel, except for length, to the full-length test from which it was derived.

Tests Used in Equating a Test Itself

For each administration, scores were available on an external anchor test (a nonoperational section of verbal material) similar in content to the SAT-V. The external anchor tests each contained equal numbers of items of the four types included in the SAT-V. The difficulties of the external anchor tests were not subject to experimentation. The mean difficulties of the external anchor tests were within about one-half a Δ_e point of the mean for SAT-V, but both the standard deviations of the Δ_e's (σ_{Δ_e}) and the mean $\underline{r}_b$'s tended to be somewhat lower for the anchor tests.

For each administration, three internal anchor tests for equating SAT-V to itself were specially constructed for the study from the pool of 85 operational items in each form. The internal anchor tests were constructed to be similar in content to SAT-V but to vary with regard to each other in mean difficulty. Internal anchor tests constructed from the April 1975 item pool each contained 10 antonym, 6 analogy, 8 sentence completion, and 10 reading comprehension items. The number of items in these respective categories from the November 1975 item pool were 10, 8, 8, and 10. For the medium-difficulty anchor tests, $\bar{\Delta}_e$, σ_{Δ_e}, and $\bar{r}_b$ were made to match the corresponding values for SAT-V as closely as possible, given the prescribed content distribution. The σ_{Δ_e}'s for the easy and hard anchor tests were, of necessity, smaller than those for SAT-V and the medium-difficulty anchor test. The item summary statistics and identification codes for the total tests and the anchor tests used in equating a test to itself are given in Table 4.

Tests Used in Equating Different Tests

For each administration, an expanded item pool consisting of the 85 SAT-V items and the 40 items in the verbal external anchor test was used for creating three total tests. The tests within each set were systematically different in average difficulty but were similar in content and equal in length. The total tests constructed from the April 1975 and the November 1975 item pools each contained 15 antonym, 13 analogy, 11 sentence completion, and 15 reading comprehension items.

For each administration a 20-item internal anchor test of medium difficulty, similar in content to SAT-V, was constructed for use in equating the special-purpose total tests. The internal anchor tests constructed from the April 1975 and the November 1975 item pools each contained 6 antonym, 5 analogy, 4 sentence completion, and 5 reading comprehension items. The item summary statistics and identification codes for the total tests and the anchor tests used in equating different tests are given in Table 4.

Samples

Two base samples were used for the study, one each from the April 1975 and the November 1975 Saturday administrations of the SAT-V. The April base sample (No. 32) was selected from those candidates who took Verbal Equating Test FM; the November base sample (No. 44), from those who took Verbal Equating Test FG. (Six base samples were used in the full study.)

Each base sample consisted of 4,731 cases, from which 5 subsamples of 1,577 cases each were created in two different ways. Two nonoverlapping subsamples ("random" samples) were selected by use of an IBM recursive random number generator. Three nonoverlapping subsamples ("dissimilar" samples) were selected by an algorithm designed to yield samples dissimilar in mean verbal ability. Two variables from the SDQ--level of educational aspiration and amount of high-school foreign language training--were used to select the dissimilar samples. These variables were known from prior information to have a high relationship with SAT-V scores.

Thus, a total of 10 subsamples of 1,577 cases each, 5 for each of the 2 base samples, were used in the study. Tables 5 and 6 give the summary statistics for the total tests and the anchor tests used in the various equating analyses.

Evaluative Procedures

Discrepancy Indices

For each raw score $\underline{x}$ there is a corresponding criterion score $\underline{t}$ and an estimated criterion score $\underline{t}'$ derived from a specific equating model. The smaller the difference $\underline{d}$ between $\underline{t}'$ and $\underline{t}$, the smaller the equating error and the more appropriate the equating model.

The standardized weighted mean square difference and squared bias were selected as the most useful summary indices for evaluating the effectiveness of the various models. The weighted mean square difference gives the greatest weight to those values of $\underline{x}$ that are most likely to occur and is consistent with what is used to represent total error in the statistical literature. The in-

Table 4
Item Summary Statistics and Identification Codes for Total Tests and Anchor Tests Used in Equating at Two Administrations

Administration	Test Description	ID	n	$\overline{\Delta}_e$	σ_{Δ_e}	$\overline{r}_b$
Equating a Test to Itself						
April 1975	Total Test					
	SAT-V	FT2XX	85	11.40	3.38	.47
	Anchor Test					
	External	FE2FM	40	11.14	3.06	.43
	Internal					
	Easy	FE2DE	34	9.39	2.90	.49
	Medium	FE2DM	34	11.40	3.26	.49
	Hard	FE2DH	34	12.96	2.49	.47
November 1975	Total Test					
	SAT-V	FT4XX	85	11.36	3.40	.48*
	Anchor Test					
	External	FE4FG	40	12.01	2.95	.43
	Internal					
	Easy	FE4DE	36	9.40	2.68	.51
	Medium	FE4DM	36	11.44	3.43	.49
	Hard	FE4DH	36	13.29	2.18	.48
Equating Different Tests						
April 1975	Total Tests					
	Easy	FT2DE	54	9.34	2.89	.46
	Medium	FT2DM	54	11.34	2.80	.46
	Hard	FT2DH	54	13.26	2.95	.44
	Anchor Test					
	Internal	FE2PA	20	11.31	3.26	.46
November 1975	Total Tests					
	Easy	FT4DE	54	9.59	2.94	.50
	Medium	FT4DM	54	11.69	2.49	.46
	Hard	FT4DH	54	13.51	2.77	.45
	Anchor Test					
	Internal	FE4PA	20	11.58	3.05	.47

*Based on 84 of the 85 items.

dices were standardized (expressed as a proportion of the criterion standard deviation) so that results could be compared across equating situations as well as across equating models.

The standardized weighted mean square difference or total error is equal to the variance of the difference plus the squared bias, that is,

$$\sum_j f_j d_j^2 / n s_t^2 = \sum_j f_j (d_j - \overline{d})^2 / n s_t^2 + \overline{d}^2 / s_t^2 , \qquad [2]$$

or Total Error = Variance of Difference + Squared Bias,

where

d_j = $(t'_j - t_j)$;
t'_j = the estimated criterion score for raw score x_j;
t_j = the criterion score for x_j;
$\bar{d}$ = $\sum_j f_j d_j / n$;
s_t = the standard deviation of the criterion scores t_j;
f_j = the frequency of x_j;
n = $\sum_j f_j$;

and the summation was over that range of x for which extrapolation was unnecessary for any of the models studied. If the ratio of the squared bias to the total error is 1, then the criterion line and the conversion line are parallel. If the difference is less than 1, then there is an interaction between the model and the criterion.

Criterion Equatings

In the case in which a test was equated to itself, the criterion for the various equatings was the test score itself. The new and old forms were treated as different tests, when in reality they were one and the same test. The ideal equating would reproduce the score on the old form exactly; that is, the conversions from raw to scaled scores would be the same for the new form as for the old form. The criterion was not so simply established in the case in which a test was equated to a different test. In these instances, it was necessary to calculate "true" conversions by equating the tests in as ideal a manner as possible.

The criterion equatings were accomplished using data from all of the cases in the two base samples (No. 32 and No. 44) from which subsamples were selected. Since all 4,731 cases in each base sample had scores on the tests being equated, it was possible to equate the scores using a single sample, an ideal equating situation. The scores could be linked directly without involving an anchor test.

Two equating methods were used to establish the two criteria against which to compare the results of the experimental equatings: equipercentile equating of estimated true scores derived from the 3-parameter logistic test model (the ICC equipercentile criterion) and equipercentile equating of observed scores (the direct equipercentile criterion). Although the criterion equatings using estimated true scores were based on ICC methodology, the method was different from that used in the experimental equatings. Nevertheless, the ICC equipercentile criterion could be biased in favor of ICC equating methods. Thus, the direct equipercentile criterion was also used. This criterion might be biased in favor of equipercentile, but not ICC, methods. For the study of the curvilinear methods reported here, both methods were appropriate, since the total tests were equal in length and yielded scores with nearly equal reliabilities. (True score equating will generally yield better conversions than observed score equating when the new and old form scores have unequal reliabilities.)

Table 5
Formula Score Means, Standard Deviations, and Correlations[a] Between Anchor Test and Total Test Scores for Base Sample No. 32 (April 1975)

		Samples[b]				
		Random		Dissimilar		
Test	No. of Items	321XX	322XX	327XX	328XX	329XX
Equating a Test to Itself						
Total Test						
FT2XX	85					
Mean		34.47	34.67		32.95	40.28
SD		15.48	15.57		14.47	14.82
Anchor Test						
FE2DE	34					
Mean		19.91	19.96		19.43	22.27
SD		6.78	6.72		6.46	6.00
$\underline{r}_{xv}$		.93	.93		.92	.92
FE2DM	34					
Mean		14.00	14.12		13.39	16.56
SD		6.98	6.98		6.58	6.66
$\underline{r}_{xv}$		.94	.94		.93	.94
FE2DH	34					
Mean		9.15	9.27		8.45	11.83
SD		7.35	7.36		6.90	7.37
$\underline{r}_{xv}$		.94	.94		.93	.95
FE2FM	40					
Mean		16.99	16.63		16.01	19.52
SD		8.06	8.06		7.78	7.86
$\underline{r}_{xv}$		.87	.87		.85	.86
Equating a Test to a Different Test						
Total Test						
FT4DE	54					
Mean		31.61		27.94		
SD		10.17		10.20		
$\underline{r}_{xv}$		.88		.86		
FT4DM	54					
Mean		21.26	21.14		21.47	
SD		11.67	11.08		10.99	
$\underline{r}_{xv}$		.90	.89		.89	
FT4DH	54					
Mean			11.89			16.17
SD			9.53			10.89
$\underline{r}_{xv}$			.89			.89
Anchor Test						
FE2PA	20					
Mean		8.46	8.42	7.25	8.02	9.72
SD		4.15	4.17	3.97	3.99	4.07

[a] Correlations are between the indicated anchor test score and total test score FT2XX in the case of a test being equated to itself, and between anchor test score FE2PA and the indicated total test score in the case of a test being equated to a different test. All anchor test scores were included in the total test score except FE2FM.

[b] The last two digits of the sample number refer to the total test and anchor test score combination used for a particular equating.

Table 6
Formula Score Means, Standard Deviations, and Correlations[a] Between Anchor Test and Total Test Scores for Base Sample No. 44 (November 1975)

Test	No. of Items	Samples[b] Random 441XX	442XX	Dissimilar 447XX	448XX	449XX
Equating a Test to Itself						
Total Test						
FT4XX	85					
Mean		36.34	35.97		36.70	42.18
SD		15.77	14.92		14.74	15.10
Anchor Test						
FE4DE	36					
Mean		21.82	21.74		22.15	24.19
SD		7.13	7.11		6.76	6.43
r_{xv}		.93	.93		.93	.92
FE4DM	36					
Mean		15.31	15.23		15.47	17.81
SD		7.13	6.85		6.75	6.89
r_{xv}		.95	.95		.94	.95
FE4DH	36					
Mean		9.19	8.98		9.12	12.02
SD		7.85	7.29		7.56	8.06
r_{xv}		.94	.93		.93	.95
FE4FG	40					
Mean		14.38	14.15		14.57	17.22
SD		8.32	8.02		7.80	8.24
r_{xv}		.87	.86		.85	.87
Equating a Test to a Different Test						
Total Test						
FT2DE	54					
Mean		31.88		28.61		
SD		10.05		10.25		
r_{xv}		.89		.87		
FT2DM	54					
Mean		21.99	22.07		20.85	
SD		10.95	10.92		10.38	
r_{xv}		.91	.90		.89	
FT2DH	54					
Mean			12.94			16.47
SD			10.21			10.45
r_{xv}			.89			.89
Anchor Test						
FE4PA	20					
Mean		8.12	7.93	6.66	8.08	9.48
SD		4.36	4.13	3.89	4.07	4.39

[a] Correlations are between the indicated anchor test score and total test score FT4XX in the case of a test being equated to itself, and between anchor test score FE4PA and the indicated total test score in the case of a test being equated to a different test. All anchor test scores were included in the total test score except FE4FG.

[b] The last two digits of the sample number refer to the total test and anchor test score combination used for a particular equating.

To determine the ICC equipercentile criterion, the 3-parameter logistic test model was applied separately to Reading (reading comprehension and sentence completion) and Vocabulary (antonym and analogy) items, so that unidimensionality did not have to be assumed across all item types. The following scores were equated:

Base Sample No. 32		Base Sample No. 44	
New Form	Old Form	New Form	Old Form
FT2DE	FT2DM	FT4DE	FT4DM
FT2DE	FT2DH	FT4DE	FT4DH
FT2DM	FT2DH	FT4DM	FT4DH

The following steps were required to accomplish the 3-parameter logistic ICC criterion equatings:

1. LOGIST was used to calculate item parameter estimates and examinee ability estimates separately for each set of relatively homogeneous items (Reading and Vocabulary).

2. For each of the three scores, for each examinee, an estimated true raw (R-W/4) score was calculated across the item types represented in each score. The true raw score (R_a) was estimated as follows:

$$R_a = \sum_i P_i(\hat{\theta}_{ia}) - [\Sigma Q_i(\hat{\theta}_{ia})]/4 \qquad [3]$$

where

$\hat{\theta}_{ia}$ is the estimated ability of examinee a on the item type represented by item i (if item i is a Reading item, then $\hat{\theta}_{ia}$ is the estimate of the examinee's reading ability, and so forth);

P_i is the probability of examinee a answering item i correctly (as calculated from the 3-parameter logistic test model);

$Q_i = 1 - P_i$; and

1/4 is the correction factor for guessing for five-choice multiple-choice items.

(Each summation was over only those items to which examinee a actually responded.)

3. For each of the three pairs of scores, the scores were directly equated by the equipercentile method. Raw and scaled score equivalents were generated for each integral score on the new form for which it was possible to establish a conversion. (The true raw scores did not usually extend over the possible score range, thus making it impossible to establish a conversion for some scores.)

RESULTS AND DISCUSSION

Comparisons of Equating Models

Tables 7 through 12 and Figures 1 through 6 summarize discrepancies between the results of the experimental equatings and the criterion equatings. The discrepancies are stated as mean square error and squared bias. To make the results comparable for different tests, the discrepancies are expressed on a scale on which the standard deviation of the criterion scores would be 100 for a standard reference group. The discrepancy indices for Tables 9 and 11 and Figures 3 and 5 were calculated in relation to criterion equatings in which the 3-parameter logistic test model was used (the ICC equipercentile criterion). The discrepancy indices for Tables 10 and 12 and Figures 4 and 6 were calculated in relation to criterion equatings in which the equipercentile equating method was used (the direct equipercentile criterion).

The test variations represented in Tables 7 through 12 are characterized in Tables 2 and 3. The statistical characteristics of the tests on which the total test scores and the anchor test scores were based are described in Table 4. The first three digits of the numeric codes in the column labeled "Sample Number" identify the subsamples on which the equatings were based (see Tables 5 and 6 for sample statistics).

The column labeled "Best Lin" identifies the linear model that had the smallest mean square error under the condition specified. The linear models considered are enumerated in Sections I (observed score models) and III (true score models) of Table 1, which also gives the identification codes for all models. The models identified in Tables 7 through 12 as Equi% (Dir), Equi% (FE), 3-Par ICC, and 1-Par ICC correspond, respectively, to Sections IIB, IIA, IVA2, and IVA1 of Table 1.

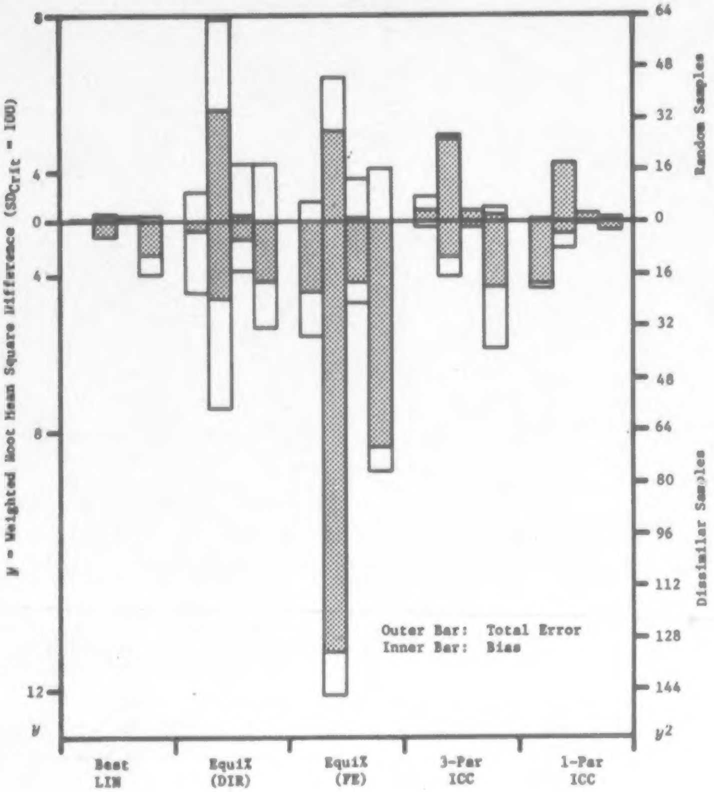
Figures 1 through 6 correspond sequentially to Tables 7 through 12. It is essential to note that the size of the scale units represented on the y-axis is different in different figures. The scale on the left side of each figure is in terms of standard deviation units and is the square root of the linear scale of mean square error shown on the right side of the figure. Contiguous bars represent the results for one model or, in the case of "Best Lin," for the best of a set of models. The results for random samples are shown in the top half of the figure; and the results for dissimilar samples, in the bottom half. The outer bars represent total error; and the inner shaded bars, bias.

The bars in each set, from left to right, appear in the successive rows in the comparable area of the corresponding table. For example, the first bar in each set in Figure 1 represents the results obtained with an internal anchor test for a pair of April 1975 samples, which were either random or of middle and high ability, respectively; the second bar, an external anchor test for the same April samples; the third bar, an internal anchor test for a pair of November 1975 samples; and the fourth bar, an external anchor test for the same November samples.

Equating a Test to Itself

Test Variations 1 and 4. The data in Table 7 (and Figure 1) show that for both similar and dissimilar samples the best linear model had the smallest amount of total error across replications, followed by the 1-parameter ICC and 3-parameter ICC models. The direct equipercentile and frequency estimation equipercentile methods had relatively more error. Surprisingly, the total error for dissimilar samples was very similar to the total error for random samples, implying, perhaps, that if the anchor test is nearly parallel to the total tests, the differences between samples is not so important. The only method showing noticeably more error for dissimilar samples was the frequency estimation equipercentile method. In this instance, bias accounted for a large proportion of the error. For dissimilar samples, the equatings through an external anchor had noticeably more error for all but the 1-parameter ICC model.

Figure 1
Comparisons of Equating Models: Equating a Test to Itself Through a Medium-Difficulty Anchor Test



Test Variation 5. For random samples the total errors for Test Variation 5 (Table 8 and Figure 2) were similar to the total errors for Test Variations 1

Table 7
Mean Square Errors and Squared Biases for Comparing Equating Models where a Test is Equated to Itself through a Medium-Difficulty Equating Test (Test Variations 1 and 4)

Test Var.	New Form Total Score[a]	New Form Sample Number	Old Form Total Score[a]	Old Form Sample Number	Equating Score[a]	Best Linear Model	Equating Model: Best Lin	Equating Model: Equi% (Dir)	Equating Model: Equi% (FE)	Equating Model: 3-Par ICC	Equating Model: 1-Par ICC
					Random Samples--Mean Square Error						
4	FT2XX	32102	FT2XX	32202	FE2DM	FE-T-A	.29	9.30	6.25	7.67	.56
1	FT2XX	32110	FT2XX	32210	FE2FM	PAC	2.34	63.20	45.02	26.83	18.40
4	FT4XX	44102	FT4XX	44202	FE4DM	LXY-IF	1.59	17.98	13.40	3.72	2.56
1	FT4XX	44110	FT4XX	44210	FE4FG	PB	1.74	17.89	16.56	4.54	1.06
					Dissimilar Samples--Mean Square Error						
4	FT2XX	32802	FT2XX	32902	FE2DM	LML-IF	.13	22.28	35.76	1.72	20.98
1	FT2XX	32810	FT2XX	32910	FE2FM	PB	4.88	58.06	146.65	17.14	8.53
4	FT4XX	44802	FT4XX	44902	FE4DM	T20-IF	.18	15.44	25.20	2.19	.85
1	FT4XX	44810	FT4XX	44910	FE4FG	LV-IR1	16.65	33.18	77.44	39.31	2.66
					Random Samples--Squared Bias[b]						
4	FT2XX	32102	FT2XX	32202	FE2DM	FE-T-A	(-) .21	(-) .32	(-) .24	(-) 3.66	(-) .56
1	FT2XX	32110	FT2XX	32210	FE2FM	PAC	(+) 1.71	(+)34.92	(+) 28.16	(+)25.76	(+)18.22
4	FT4XX	44102	FT4XX	44202	FE4DM	LXY-IF	(-) 1.59	(-) 2.01	(-) 1.08	(-) 3.45	(-) 2.51
1	FT4XX	44110	FT4XX	44210	FE4FG	PB	(+) .65	(+) .24	(+) .32	(-) 2.38	(-) .39
					Dissimilar Samples--Squared Bias[b]						
4	FT2XX	32802	FT2XX	32902	FE2DM	LML-IF	(-) .01	(+) 3.02	(+) 22.00	(-) .06	(-)19.51
1	FT2XX	32810	FT2XX	32910	FE2FM	PB	(-) 4.83	(+)24.08	(+)133.29	(+)11.14	(+) 4.03
4	FT4XX	44802	FT4XX	44902	FE4DM	T20-IF	(+) .18	(+) 5.82	(+) 18.83	(+) 1.94	(-) .15
1	FT4XX	44810	FT4XX	44910	FE4FG	LV-IR1	(+)10.48	(+)18.65	(+) 69.88	(+)20.36	(+) 2.53

[a] Scores were based on the following numbers of items: FT2XX and FT4XX, 85; FE2FM and FE4FG, 40; FE2DM, 34; FE4DM, 36.

[b] The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 8

Mean Square Errors and Squared Biases for Comparing Equating Models where a Test is Equated to Itself through an Easy or a Hard Equating Test (Test Variation 5)

New Form		Old Form			Best	Equating Model				
Total Score[a]	Sample Number	Total Score[a]	Sample Number	Equating Score[a]	Linear Model	Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
				Random Samples--Mean Square Error						
FT2XX	32101	FT2XX	32201	FE2DE	LXY-IR1	.22	11.90	6.05	3.28	.34
FT2XX	32103	FT2XX	32203	FE2DH	T1	.31	10.11	6.15	8.76	.20
FT4XX	44101	FT4XX	44201	FE4DE	PAC	6.10	33.87	27.98	11.63	1.37
FT4XX	44103	FT4XX	44203	FE4DH	FE-T-F	.45	10.18	8.35	5.20	.11
				Dissimilar Samples--Mean Square Error						
FT2XX	32801	FT2XX	32901	FE2DE	LXY-IF	30.14	29.70	53.73	2.34	2.69
FT2XX	32803	FT2XX	32903	FE2DH	LV-IR1	22.85	17.39	39.94	1.28	2.07
FT4XX	44801	FT4XX	44901	FE4DE	FE-T-A	59.14	36.36	61.94	15.37	2.40
FT4XX	44803	FT4XX	44903	FE4DH	FE-O	1.19	19.10	9.18	2.72	5.34
				Random Samples--Squared Bias[b]						
FT2XX	32101	FT2XX	32201	FE2DE	LXY-IR1	(+) .21	(+) .31	(+) .40	(+) .26	(+) .27
FT2XX	32103	FT2XX	32203	FE2DH	T1	(-) .08	(-) .41	(-) .21	(-) 3.63	(-) .16
FT4XX	44101	FT4XX	44201	FE4DE	PAC	(-) 6.07	(-) 1.36	(-) .88	(-) 2.53	(-) .82
FT4XX	44103	FT4XX	44203	FE4DH	FE-T-F	(-) .45	(-) .64	(-) .37	(-) 1.07	(-) .06
				Dissimilar Samples--Squared Bias[b]						
FT2XX	32801	FT2XX	32901	FE2DE	LXY-IF	(-) 1.27	(+) 11.66	(+) 41.91	(+) .17	(-) .65
FT2XX	32803	FT2XX	32903	FE2DH	LV-IR1	(+) 3.04	(+) 5.44	(+) 28.10	(+) .60	(-) .59
FT4XX	44801	FT4XX	44901	FE4DE	FE-T-A	(+) 58.19	(+) 27.93	(+) 56.22	(+) 14.70	(+) 2.33
FT4XX	44803	FT4XX	44903	FE4DH	FE-O	(+) .90	(-) 1.78	(+) .87	(-) 1.16	(-) 4.88

---

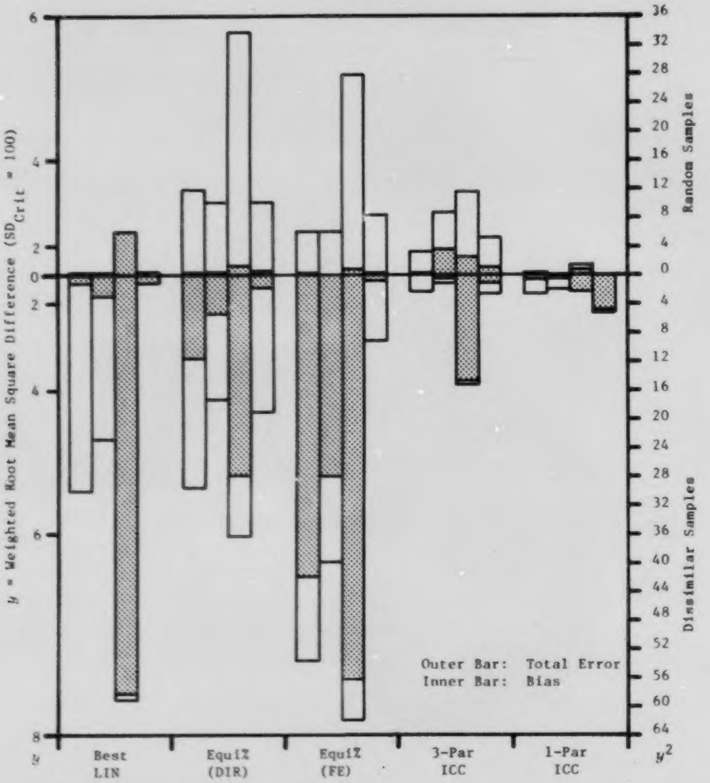
[a] Scores were based on the following numbers of items: FT2XX and FT4XX, 85; FE2FM and FE4FG, 40; FE2DM, 34; FE4DM, 36.

[b] The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

and 4. The fact that the total error for the best linear model was small for random samples suggests that the curvilinear relation of the anchor test and the total test had little effect on the equating.

The total error for the best linear model was substantially larger when dissimilar samples were used. The ICC models were noticeably superior to the other models in this case, suggesting that these models are relatively robust when the anchor test is different in difficulty from the total test and the samples differ in ability.

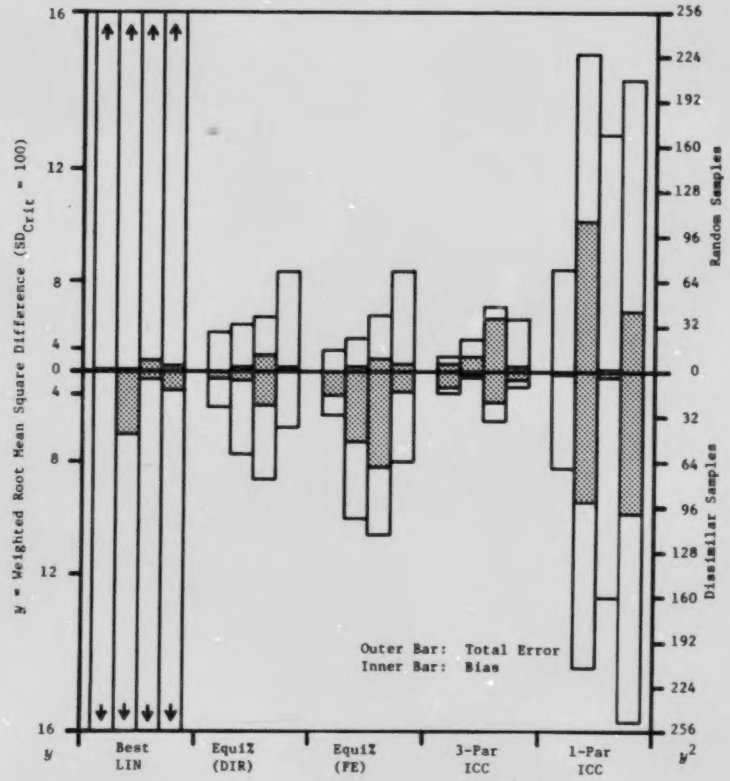
Figure 2
Comparisons of Equating Models: Equating a Test to Itself Through an Easy or Difficult Anchor Test



Equating a Test to a Different Test

Test Variations 8 and 9. The introduction of some curvilinearity in the relation between the scores of the tests being equated (due to differences in difficulty) resulted, as expected, in a very large total error for the best linear model (see Tables 9 through 12 and Figures 3 to 6). For the most part, sample variation seemed to have little effect on total error for the curvilinear models. All of the curvilinear models had noticeably less error than the best linear model, but the 1-parameter ICC model had substantially more error than the other curvilinear models. For most equatings, the model with the smallest total error was the 3-parameter ICC model. For the most part, when greater differences in total test difficulty were introduced (greater curvilinearity), the total error tended to increase substantially for all models, becoming exceedingly large for the best linear model.

Figure 3
Comparisons of Equating Models: Equating a Test to a Different Test Using Easy and Medium or Medium and Difficult Tests (ICC Equipercentile Criterion)



Criterion Bias

Equating a Test to a Different Test

The total error and bias for the two criteria for Test Variations 8 and 9 can be compared in Figures 3 and 4 (Tables 9 and 10) and in Figures 5 and 6 (Tables 11 and 12). It may be noted that for random samples the experimental equipercentile equatings had less error than the 3-parameter ICC model when the equipercentile observed-score criterion was used, suggesting that the equipercentile criterion equatings based on true scores may be biased in favor of the 3-parameter ICC model, just as the equipercentile criterion equatings based on observed scores seem to be biased in favor of the equipercentile models. Inter-

Table 9

Mean Square Errors and Squared Biases (ICC Equipercentile Criterion) for Comparing Equating Models where a Test is Equated to a Different, More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 8)

New Form		Old Form			Best	Equating Model				
Total Score[a]	Sample Number	Total Score[a]	Sample Number	Equating Score[a]	Linear Model	Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
					Random Samples--Mean Square Error					
FT2DE	32115	FT2DM	32217	FE2PA	TML-IA[b]	411.28	27.77	15.37	10.82	72.93
FT2DM	32117	FT2DH	32216	FE2PA	LV-IF	547.56	33.18	23.91	22.94	225.90
FT4DE	44115	FT4DM	44217	FE4PA	FE-O	656.90	38.94	40.32	45.97	169.00
FT4DM	44117	FT4DH	44216	FE4PA	LV-IF	469.59	71.23	71.57	37.09	207.94
					Dissimilar Samples--Mean Square Error					
FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	417.79	25.50	31.36	15.84	68.89
FT2DM	32817	FT2DH	32916	FE2PA	FE-T-A	608.61	59.14	105.06	4.49	211.12
FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	679.12	77.44	116.21	35.40	160.53
FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	482.68	40.07	64.80	11.36	249.32
					Random Samples--Squared Bias[c]					
FT2DE	32115	FT2DM	32217	FE2PA	TML-IA[b]	(+) 1.47	(+) .77	(+) .37	(-) 5.79	(+) .02
FT2DM	32117	FT2DH	32216	FE2PA	LV-IF	(-) 1.85	(-) 3.78	(-) 3.86	(-)10.95	(-)107.26
FT4DE	44115	FT4DM	44217	FE4PA	FE-O	(+) 8.43	(+)12.08	(+) 9.49	(+)38.10	(+) 2.21
FT4DM	44117	FT4DH	44216	FE4PA	LV-IF	(-) 4.38	(-) 3.24	(-) 5.34	(-) 3.85	(-) 43.03
					Dissimilar Samples--Squared Bias[c]					
FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	(-) .36	(+) 5.36	(+) 17.27	(-)11.69	(+) 2.14
FT2DM	32817	FT2DH	32916	FE2PA	FE-T-A	(+)45.20	(+) 6.74	(-) 50.63	(-) 2.29	(-) 93.00
FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	(-) 6.11	(+)24.56	(+) 68.33	(+)22.42	(+) 4.41
FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	(-)13.45	(+) 1.06	(+) 15.01	(-) 6.29	(-)101.20

[a] Scores were based on the following numbers of items: FT2DE, FT2DM, FT2DH, FT4DE, FT4DM, and FT4DH, 54; FE2PA and FE4PA, 20.

[b] Also TML-IF, TML-IR1, LUR-IA, LXY-IA, LV-IA, LML-IA, and PB.

[c] The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 10

Mean Square Errors and Squared Biases (Direct Equipercentile Criterion) for Comparing Equating Models where a Test is Equated to a Different, More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 8)

New Form		Old Form			Best	Equating Model				
Total Score[a]	Sample Number	Total Score[a]	Sample Number	Equating Score[a]	Linear Model	Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
				Random Samples--Mean Square Error						
FT2DE	32115	FT2DM	32217	FE2PA	LML-IF	337.82	13.54	7.45	23.04	64.48
FT2DM	32117	FT2DH	32216	FE2PA	LV-IR1	361.00	8.01	9.92	34.57	102.62
FT4DE	44115	FT4DM	44217	FE4PA	FE-0	528.08	21.72	15.76	48.72	129.73
FT4DM	44117	FT4DH	44216	FE4PA	PD	273.90	13.54	11.90	36.36	100.00
				Dissimilar Samples--Mean Square Error						
FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	344.10	15.92	20.52	29.81	58.06
FT2DM	32817	FT2DH	32916	FE2PA	LV-IR1	431.39	43.30	114.28	19.62	89.11
FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	543.82	45.83	85.01	31.47	118.81
FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	275.23	25.10	44.22	34.57	129.73
				Random Samples--Squared Bias[b]						
FT2DE	32115	FT2DM	32217	FE2PA	LML-IF	(+) .80	(+) .28	(+) .07	(-) 7.68	(-) .05
FT2DM	32117	FT2DH	32216	FE2PA	LV-IR1	(+)2.53	(+) 1.06	(+) 1.02	(-) .11	(-)54.97
FT4DE	44115	FT4DM	44217	FE4PA	FE-0	(+)8.99	(+)12.84	(+) 10.12	(+)39.86	(+) 2.47
FT4DM	44117	FT4DH	44216	FE4PA	PD	(-) .01	(+) .36	(+) .01	(+) .20	(-)16.88
				Dissimilar Samples--Squared Bias[b]						
FT2DE	32715	FT2DM	32817	FE2PA	LV-IR1	(-) .91	(+) 3.92	(+) 14.72	(-)14.36	(+) 1.26
FT2DM	32817	FT2DH	32916	FE2PA	LV-IR1	(-)2.59	(+)31.30	(+)102.63	(+) 2.15	(-)44.89
FT4DE	44715	FT4DM	44817	FE4PA	LV-IF	(-)5.97	(+)25.78	(+) 71.08	(+)23.59	(+) 4.77
FT4DM	44817	FT4DH	44916	FE4PA	LV-IF	(-)1.56	(+)11.58	(+) 38.65	(-) .01	(-)57.31

[a] Scores were based on the following numbers of items: FT2DE, FT2DM, FT2DH, FT4DE, FT4DM, and FT4DH, 54; FE2PA and FE4PA, 20.

[b] The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 11

Mean Square Errors and Squared Biases (ICC Equipercentile Criterion) for Comparing Equating Models where a Test is Equated to a Different, Much More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 9)

New Form		Old Form			Best	Equating Model				
Total Score[a]	Sample Number	Total Score[a]	Sample Number	Equating Score[a]	Linear Model	Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
				Random Samples--Mean Square Error						
FT2DE	32115	FT2DH	32216	FE2PA	PAC	1,710.65	84.46	70.22	61.47	412.09
FT4DE	44115	FT4DH	44216	FE4PA	LXY-IR1	1,921.95	131.56	137.36	23.23	547.09
				Dissimilar Samples--Mean Square Error						
FT2DE	32715	FT2DH	32916	FE2PA	PD	1,704.04	88.55	178.49	30.80	327.61
FT4DE	44715	FT4DH	44916	FE4PA	PD	1,927.21	166.15	241.80	30.58	495.95
				Random Samples--Squared Bias[b]						
FT2DE	32115	FT2DH	32216	FE2PA	PAC	(-) 12.94	(-)12.39	(-) 13.41	(-)43.24	(-)188.18
FT4DE	44115	FT4DH	44216	FE4PA	LXY-IR1	(-) 3.17	(-) 5.28	(-) 7.55	(-) .43	(-)137.14
				Dissimilar Samples--Squared Bias[b]						
FT2DE	32715	FT2DH	32916	FE2PA	PD	(+) .37	(+)17.46	(+)101.59	(-)24.54	(-) 97.89
FT4DE	44715	FT4DH	44916	FE4PA	PD	(+) .38	(+)19.12	(+)117.57	(+) .85	(-) 94.08

[a] Scores were based on the following numbers of items: FT2DE, FT2DM, FT4DE, FT4DH, 54; FE2PA and FE4PM, 20.

[b] The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

Table 12

Mean Square Errors and Squared Biases (Direct Equipercentile Criterion) for Comparing Equating Models where a Test is Equated to a Different, Much More Difficult Test through a Medium-Difficulty Equating Test (Test Variation 9)

New Form		Old Form			Best	Equating Model				
Total Score[a]	Sample Number	Total Score[a]	Sample Number	Equating Score[a]	Linear Model	Best Lin	Equi% (Dir)	Equi% (FE)	3-Par ICC	1-Par ICC
				Random Samples--Mean Square Error						
FT2DE	32115	FT2DH	32216	FE2PA	PAC	1,253.16	10.76	12.89	101.40	247.43
FT4DE	44115	FT4DH	44216	FE4PA	LXY-IF	1,326.42	16.89	16.08	87.24	302.76
				Dissimilar Samples--Mean Square Error						
FT2DE	32715	FT2DH	32916	FE2PA	TML-IA[b]	1,254.58	69.39	191.27	49.98	171.61
FT4DE	44715	FT4DH	44916	FE4PA	T20-IR1	1,334.44	85.75	215.50	53.29	254.40
				Random Samples--Squared Bias[c]						
FT2DE	32115	FT2DH	32216	FE2PA	PAC	(-) .15	(-) .09	(-) .19	(-)11.71	(-)114.61
FT4DE	44115	FT4DH	44216	FE4PA	LXY-IF	(+) 1.42	(+) .43	(+) .04	(+)13.37	(-) 79.15
				Dissimilar Samples--Squared Bias[c]						
FT2DE	32715	FT2DH	32916	FE2PA	TML-IA[b]	(-) 24.93	(+)57.04	(+)184.43	(-) 3.11	(-) 46.30
FT4DE	44715	FT4DH	44916	FE4PA	T20-IR1	(-) 7.53	(+)55.03	(+)195.79	(+)15.41	(-) 47.08

[a] Scores were based on the following numbers of items: FT2DE, FT2DH, FT4DE, and FT4DM, 54; FE2PA and FE4PM, 20.

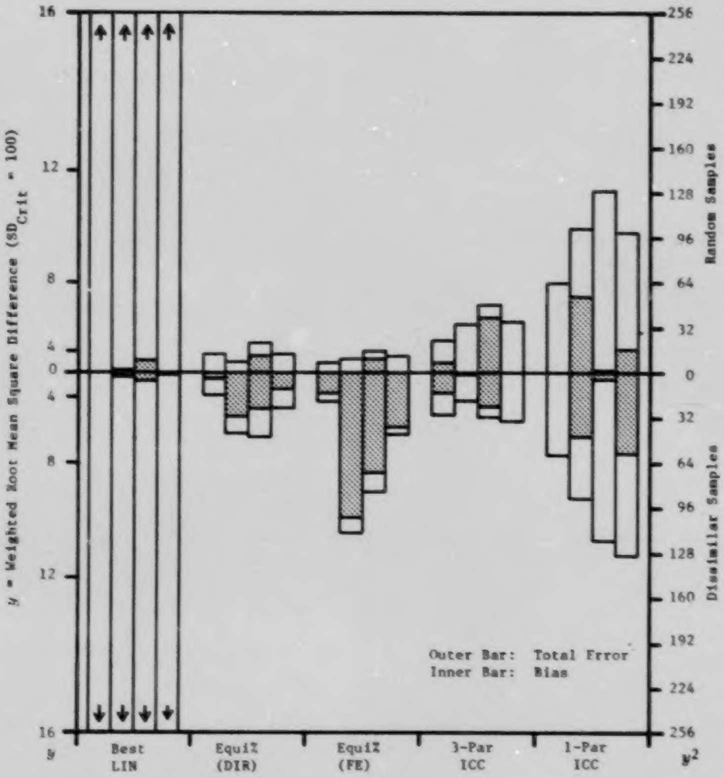
[b] Also TML-IF, TML-IR1, LUR-IA, LXY-IA, LV-IA, LML-IA, and PB.

[c] The sign in parentheses indicates the direction of the bias; a minus sign indicates that the equating model mean was lower than the criterion mean; a plus sign, that the mean was higher.

estingly enough, the 3-parameter ICC model has less total error than the 1-parameter ICC model regardless of which criterion was used, although the difference between the models was less for the direct equipercentile criterion.

The criterion bias was less obvious in the case of dissimilar samples; in fact, the rank ordering of the models in terms of total error was not affected by the choice of criterion. The 3-parameter ICC model had the smallest error under both criteria, but the size of the error decreased for the equipercentile and 1-parameter ICC models and increased for the 3-parameter ICC model when the direct equipercentile criterion was used.

Figure 4
Comparisons of Equating Models: Equating a Test to a Different Test Using Easy and Medium or Medium and Difficult Total Tests (Direct Equipercentile Criterion)



Equating a Test to Itself

The criterion would seem to be well established when a test is equated to itself. However, it is possible that the criterion in this case is biased in favor of a model that comes closest to fixing all of the item parameters at the same values, in this case, the 1-parameter ICC model.

It is easily seen that if the a's, b's, and c's in the 3-parameter logistic test model are fixed at some constant values, the true raw scores for the old and new forms corresponding to a given ability level have to be the same, since the probability of getting a particular item correct is then a function of only the item parameters and ability.

In the 1-parameter logistic test model, c is fixed at 0 and a at 1 for all items. Since only the b's are estimated, it might be expected that the 1-parameter ICC model is more likely than the 3-parameter ICC model to yield the appropriate conversions. Is it of any consequence, then, if it is found that the 1-parameter ICC model seems to be superior to the 3-parameter ICC model in equating a test to itself? What is really desired is to know which of the models works best when a test is equated to a different test, particularly when a test is equated to a parallel test. It would seem, then, that there may be a natural bias in favor of the 1-parameter ICC model when a test is equated to itself, but this probably does not affect the other models.

Figure 5
Comparisons of Equating Models: Equating a Test to a Different Test Using Easy and Difficult Total Test (ICC Equipercentile Criterion)

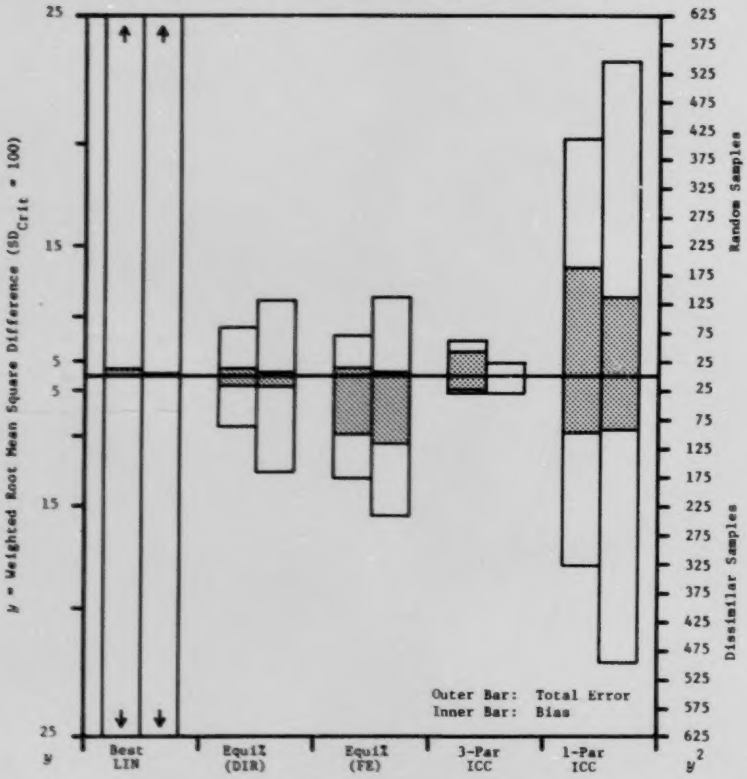
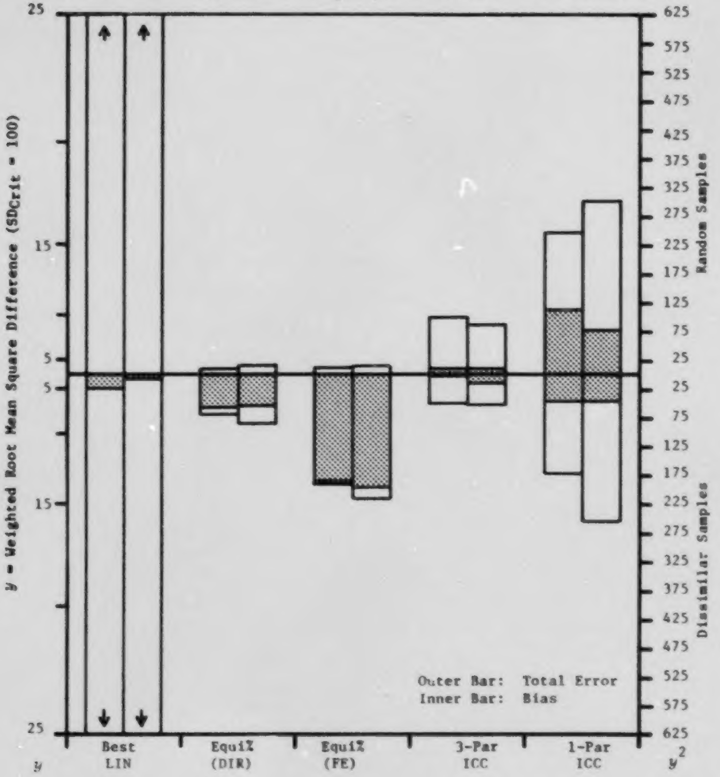


Figure 6
Comparisons of Equating Models: Equating a Test to a Different Test Using Easy and Difficult Total Test (Direct Equipercentile Criterion)



CONCLUSIONS

Several conclusions can be drawn from the study. They should be considered tentative because of possible criterion bias, which has been only partially controlled; because it was not possible to study all test variations in the full design, in particular, the variations in which parallel, but not identical, tests were equated; and because the results showed occasional inconsistencies that have not yet been explained:

1. When a test is equated to itself (or, to generalize, to a test like itself) through a parallel anchor test, a linear model yields very good results regardless of the type of samples used. However, whether any particular linear model consistently gives satisfactory results requires further study. The best linear model for a particular test variation and type of sample was used here.

2. Curvilinear models give results nearly comparable to those of the best linear model when a test is equated to a test like itself through an internal anchor. When an external anchor is used, of the curvilinear models the ICC models (particularly the 1-parameter model) give rela-

tively better results. However, the criterion may be biased in favor of the 1-parameter model.

3. The types of samples have a relatively small and unsystematic effect on the quality of the equating results if the anchor test is similar in content and in difficulty to the total tests. The one exception is the frequency estimation procedure, which seems not to perform well when an external anchor and dissimilar samples are used.

4. The equatings involving an internal anchor have less total error than comparable equatings with an external anchor. Whether or not this inconsistency would obtain when a test is equated to a test of different difficulty needs to be studied. The possibility that the criterion is biased in favor of an internal anchor also needs investigation. However, it may simply be due to the fact that the external anchor tests were not quite as similar to the total tests in content and statistical characteristics as were the internal anchor tests.

5. When a test is equated to a test like itself through an easy or hard anchor test with random samples, all of the models have a small mean square error. When dissimilar samples are used, however, the ICC models give clearly superior results.

6. When total tests differ considerably in difficulty, linear models yield unsatisfactory results in that the mean square error becomes very large; but they tend to yield better estimates of the mean than the curvilinear models. The 1-parameter ICC model and the frequency estimation method also give unacceptable results in many instances. This result is consistent with the Slinde and Linn (1978) findings that the Rasch model yielded poor results for vertical equating.

7. The 3-parameter ICC model is the best equating model when total tests of unequal difficulty are equated through a medium-difficulty anchor test with dissimilar samples. It would appear that the 3-parameter model is the equating model most likely to yield acceptable results under unusual or extreme conditions.

This study, though comprehensive in its coverage, is limited in that SAT-V items made up the item pool. These items are known to be relatively homogeneous and somewhat difficult for the current test-taking population. Similar studies are needed in situations where the content of the anchor test is allowed to depart by varying degrees from the content of the total tests and where the test-taking population is from a different age group.

REFERENCES

Angoff, W. H. Test reliability and effective test length. Psychometrika, 1953, 18, 1-14.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Angoff, W. H., & Dyer, H. S. The admissions testing program. In W. H. Angoff (Ed.), The College Board admissions testing program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board, 1971.

Feldt, L. S. Estimation of the reliability of a test divided into two parts of unequal length. Psychometrika, 1975, 40, 557-561.

Levine, R. S. Equating the score scales of alternate forms administered to samples of different ability (ETS RB-55-23). Princeton, NJ: Educational Testing Service, 1955.

Lord, F. M. A survey of equating methods based on item characteristic curve theory (ETS RB-75-13). Princeton, NJ: Educational Testing Service, 1975.

Lord, F. M. Personal communication, February 24, 1976.

Potthoff, R. F. Equating of grades or scores on the basis of a common battery of measurements. In P. R. Krishnaiah, Multivariate analysis. New York: Academic Press, 1966.

Rentz, R. R., & Bashaw, W. L. Equating reading tests with the Rasch model (Vols. 1 & 2). Athens: University of Georgia, College of Education, Educational Research Laboratory, September 1975.

Slinde, J. A., & Linn, R. L. Vertically equated tests: Fact or phantom? Journal of Educational Measurement, 1977, 14, 23-32.

Slinde, J. A., & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Tucker, L. R. Personal communication, August 7, 1974.

Wood, R. L., & Lord, F. M. A user's guide to LOGIST (ETS RM-76-4). Princeton, NJ: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (ETS RM-76-6). Princeton, NJ: Educational Testing Service, 1976.

ACKNOWLEDGMENTS

The authors are listed alphabetically. This study was funded jointly by the College Entrance Examination Board and Educational Testing Service. The authors gratefully acknowledge the advice and assistance provided by William H. Angoff, Rebecca Bullock, Samuel A. Livingston, Frederic M. Lord, Ruth C. Mroczka, Joanne S. Peng, William B. Schrader, Ledyard R. Tucker, Marilyn S. Wingersky, and R. L. Wood.

THE EFFECTS OF CONTEXT ON LATENT-TRAIT MODEL ITEM PARAMETER AND TRAIT ESTIMATES

WENDY M. YEN
CTB/MCGRAW HILL

Latent trait models hold the promise of being particularly useful in the development of test item pools. Items for an item pool can be accumulated by administering different sets of items to different groups of examinees; all the items' parameters can then be linked to the same scale through a common subset of items called anchor or linking items. After an item pool is created, different examinees can take different items from the pool, and all examinees' trait estimates should be on a common scale. An examinee's trait estimate should not be systematically affected by the choice of items, although unsystematic effects, as reflected in the standard error of measurement, can occur.

The successful use of latent trait models in the development and use of item pools relies on (1) the lack of systematic effects on item parameter estimates when they are obtained in different contexts with different examinees and (2) the lack of systematic effects on trait estimates when they are obtained with different items. Previous research on the invariance of item parameter estimates has typically examined the effects of changing samples on a fixed group of items and has not examined variations in the identities and arrangements of the items. The present research examines such variations, as well as the effects of changes in items on trait estimates. The effects of the choice of latent trait model on parameter stability and trait equating are also examined.

Method

Construction of Test Booklets

Items were chosen from the California Achievement Tests, Forms C and D (1977), Level 14, Reading Comprehension (Reading--80 items), and Level 16, Mathematics Concepts and Applications (Mathematics--90 items). The Reading items all had four-answer choices and the Mathematics items all had five-answer choices. Preliminary analyses were performed on data for students who took both test forms (N = 294 for Reading, N = 379 for Mathematics). Chi-square goodness-of-fit statistics were used to evaluate the items in terms of their fit to a 2-parameter logistic model; item difficulties and discriminations were reviewed to identify items with extreme difficulties or discriminations.

Using these results, five sets of items were created in each content area.

The first set of items (Set A) had a range of difficulties and relatively good model fit; these items were used as anchor items for linking item parameters obtained in different booklets. The second and third sets of items (Sets V and W) had relatively poor fit and, in some cases, extreme difficulties or low discriminations; these items were included to alter item contexts. The fourth and fifth sets of items (Sets X and Y) had relatively good model fit and discrimination and nonextreme difficulties, and were the items of major interest. Table 1 contains the number of items chosen for each set for each content area.

Table 1
Number of Items in Each Set

Set	Number of Items	
	Reading	Mathematics
A	10	11
V	10	11
W	10	11
X	20	22
Y	20	22

Using these sets of items, seven booklets were created, as described in Table 2. The items in the different sets were intermingled within the booklets, and the sequences of items were varied over booklets. Because of the connection of Reading items to passages and the connection of some Mathematics items to graphs, there was necessarily some similarity over booklets in the local contexts for some items. The sequence of answer choices for an item was held constant over all booklets.

Table 2
Composition of Test Booklets

Booklet	Sets	Number of Items	
		Reading	Mathematics
1	X+Y	40	44
2	A+V+X	40	44
3	A+W+Y	40	44
4	A+X	30	33
5	A+Y	30	33
6	A+X	30	33
7	A+Y	30	33

To indicate the degree of similarity of the sequence of X and Y items across booklets, the sequential positions of the X items and of the Y items were determined. Spearman rank-order correlations between items' positions in Booklet 1 and their positions in the other booklets are contained in Table 3.

Table 3
Spearman Rank-Order Correlations Between the Position of X or Y Items in Booklet 1 and the Position of Those Items in Booklets 2 to 7

Booklet	Set	Correlation with Booklet 1 Reading	 Mathematics
2	X	.90	.05
3	Y	.56	.40
4	X	.59	-.06
5	Y	.70	.62
6	X	.10	-.06
7	Y	.30	.03

Note. Each correlation is based on 20 items for Reading and 22 items for Mathematics.

Test Administration

Students were tested in Grade 4 for Reading and in Grade 6 for Mathematics. Time limits for test administration were adjusted for the length of the booklet and were made comparable (on a time-per-item basis) to those given in the *California Achievement Tests Examiner's Manual, Levels 14-19, Forms C and D* (1977). For the first testing each student took one of the seven booklets. Booklets 2 and 3 (as well as Booklets 4 and 5 and Booklets 6 and 7) were administered to students in the same classrooms on an alternate-seat basis. (This alternate-seat testing was done for a study of equipercentile equating that will not be reported here.) Two weeks later, all students took Booklet 1. The number of examinees with usable answer sheets for both first and second testings appear in Table 4.

Table 4
Number of Examinees with Usable Answer Sheets for Both First and Second Testings

First-Testing Booklet*	Reading (Grade 4)	Mathematics (Grade 6)
1	470	450
2	225	228
3	216	232
4	193	230
5	198	219
6	186	232
7	190	221

Note. Booklets 2 and 3, 4 and 5, and 6 and 7 were administered on an alternate-seat basis.
*All second testings used Booklet 1.

Latent Trait Models

Two latent trait models were used--the 3-parameter logistic model and the 1-parameter (Rasch) logistic model (see Allen & Yen, 1979, for a further description of these models). For the 3-parameter logistic model, the item characteristic function for the $\underline{i}^{th}$ item and the $\underline{k}^{th}$ examinee is

$$p_i(\theta_k) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_k - b_i)}} , \qquad [1]$$

where θ_k is the latent trait value for examinee $\underline{k}$ and $\underline{a_i}$, $\underline{b_i}$, and $\underline{c_i}$ are the discriminating power, difficulty, and lower asymptote, respectively, for item $\underline{i}$. The item characteristic function for the 1-parameter model is

$$p_i(\theta_k) = \frac{1}{1 + e^{-1.7a(\theta_k - b_i)}} , \qquad [2]$$

where $\underline{a}$ is the item discrimination power common to all the items.

To obtain latent trait and item parameter estimates, examinees' item response vectors were analyzed using the LOGIST computer program provided by Wood, Wingersky, and Lord (1976). Trait estimates were not obtained for examinees who had zero or perfect scores, and using a default option of the Wood et al. (1976) program, trait estimates also were not obtained for examinees who did not answer at least a third of the items being scaled.

Item Linking

The item parameters for the X and Y item subsets were placed on the same scale using different procedures for Booklet 1 and Booklets 2 to 7. For Booklet 1, examinees took both X and Y items; item response vectors for the two sets were analyzed together, and their item parameters were automatically placed on the same scale. This procedure was followed whether Booklet 1 was administered in the first testing or in the second testing. In some analyses, those examinees who took Booklet 1 at the first testing were divided into two approximately equal-sized groups, and item parameters were obtained separately in the two groups. For convenience, these groups were labeled Booklets 1A and 1B, even though the 1A and 1B test booklets were the same; "A" and "B" merely indicated different samples of examinees.

For Booklets 2 to 7, booklets were linked in pairs: 2 and 3, 4 and 5, 6 and 7. For example, the responses to all the items in Booklets 2 and 3 were pooled and analyzed jointly. This was done by treating the items in the A, V, W, X, and Y sets as if they were all contained in one test booklet. Examinees' responses were used for all the items they completed, and they were given a "not reached" code for every item they did not take. Examinees who took Booklet 2 were given "not reached" for items in Sets W and Y, and those who took Booklet 3 were given "not reached" for items in Sets V and X. Using the LOGIST program, an examinee's trait value was based only on the items the examinee actually took; not reached items were ignored. Similarly, an item's parameters were estimated using only the responses of examinees who actually completed the item. This joint analysis of Booklets 2 and 3 placed the parameters for the A, V, W, X, and Y items on the same scale.

For some analyses, responses for Booklets 2 to 7 were pooled and jointly analyzed. Thus, responses to Booklets 2, 4, and 6 entered into the estimation of the item parameters for the X subset, responses to Booklets 3, 5, and 7 entered into the estimation of the item parameters for the Y subset; and responses to all six booklets entered into the estimation of the item parameters for the A subset of items.

Analysis of Context Effects

To examine context effects on the means and standard deviations of the item parameters from the first testing, it was necessary to place the item parameters estimated from different samples and booklets on the same scale. To do this, the item parameters obtained from the pooling of Booklets 2, 4, 6 and 3, 5, 7 were scaled so that the corresponding trait estimates had a mean of 0 and a standard deviation of 1, and mean item difficulty and mean item discriminations were obtained from the Set A items. The item parameters from the other pairs of first-testing booklets that contained Set A items (i.e., Booklets 2 and 3, 4 and 5, 6 and 7) were linearly transformed so that their Set A mean item difficulties and mean discriminating powers equaled the Set A means obtained from the pooled 2, 4, 6 and 3, 5, 7 booklets. This transformation theoretically placed all the first-testing item parameters on the same scale--a scale that produced trait estimates with a mean of 0 and a standard deviation of 1 for examinees who took Booklets 2 to 7. The X and Y item parameters could then be compared across booklets to examine systematic context effects. (Note that this comparison would not be made for Booklet 1, which did not contain Set A items.)

The square root of the mean square difference (RMSD) between two sets of estimated statistics (e.g., item parameters or trait values) was found as

$$RMSD_{m,m'} = \left[\frac{1}{n} \sum_{i=1}^{n} (z_{im} - z_{im'})^2 \right]^{\frac{1}{2}} , \qquad [3]$$

where z_{im} and $z_{im'}$ represent statistics in sets m and m' and there are n statistics being compared. The RMSD also can be expressed as

$$RMSD_{m,\, m'} = \left[s_m^2 + s_{m'}^2 + (\bar{z}_m - \bar{z}_{m'})^2 - 2 r_{mm'} s_m s_{m'} \right]^{\frac{1}{2}} , \qquad [4]$$

where

s_m and $s_{m'}$ are the standard deviations of the statistics in the two sets,
$\bar{z}_m$ and $\bar{z}_{m'}$ are the means of the statistics in the two sets, and
$r_{mm'}$ is the correlation between the two sets of statistics.

Chi-square goodness-of-fit statistics were calculated in the following fashion. Examinees were rank ordered on the basis of their trait estimates and then divided into 10 cells with approximately equal numbers of examinees in each cell. The chi-square for an item was

$$\chi^2_i = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})} , \qquad [5]$$

where

N_j was the number of examinees in cell _j_,
O_{ij} was the observed proportion of examinees in cell _j_ that passed item _i_, and
E_{ij} was the proportion of examinees in cell _j_ expected to pass item _i_.

$$E_{ij} = \frac{1}{N_j} \sum_{k \varepsilon \text{cell } j} \hat{p}_{i(\theta k)}, \qquad [6]$$

where $\hat{p}_i(\theta_k)$ was the item characteristic function evaluated using the trait estimate for examinee _k_ and the estimated item parameters for item _i_. When item parameters were estimated from the data on which the chi-square is based, this chi-square statistic had 10 - 3 = 7 degrees of freedom for the 3-parameter model and 10 - 1 = 9 degrees of freedom for the 1-parameter model.[1]

Results

Item Parameter Estimates

Table 5 contains the number of examinees whose responses entered into the first-testing item parameter estimations for the various booklets. These sample sizes usually were slightly smaller than those in Table 4 because examinees were not used if they did not answer at least a third of the items or if they had zero or perfect scores.

First testing. For the 3-parameter model the lower asymptotes, _c_, had homogeneous values centered at .20 for Reading and .15 for Mathematics. Correlations of the X + Y difficulty and discrimination parameters estimated in the different first-testing booklets are contained in Table 6.[2] Recall that Booklets 1A, 1B, and 1 had the same context and differed only in terms of the samples and sample sizes used to estimate their parameters. The correlations between Book-

---

[1] A simulation study found that when 40 item responses for 500 pseudo-examinees were generated for a model and these responses were used to estimate traits and item parameters for that model, the resulting chi-squares had expectations approximately equal to their degrees of freedom.

[2] Note that these correlations were a function of the particular items being correlated, and the correlations cannot be meaningfully compared across content areas.

Table 5
Sample Sizes Involved in Parameter
Estimations for the First Testing

Booklets	Sample Size Reading	 Mathematics
1A	232	225
1B	228	224
1	460	449
2&3	223 + 214	228 + 232
4&5	187 + 195	228 + 218
6&7	183 + 184	228 + 218
2,4,6&3,5,7	593 + 593	684 + 668

Note. Some of these sample sizes are slightly smaller than the corresponding sample sizes in Table 3 because examinees were excluded if they had zero or perfect scores or if they did not answer at least a third of the items.

lets 1A and 1B can be compared to the correlations among Booklets 2 and 3, 4 and 5, and 6 and 7 to examine the degree to which changes in context affect the stability of item parameters for sample sizes of about 200. It is clear from the data in Table 6 that a change in context substantially decreased the stability of all the item parameter estimates.

First versus second testing. Another comparison of the correlations between item parameters was made. Item parameters were obtained for the X + Y items using all the second-testing data. (Recall that all the second testings used the same booklet, Booklet 1.) These parameter estimates were based on relatively large sample sizes and therefore had fairly small standard errors. Correlations between these second-testing parameter estimates and the first-testing parameter estimates are contained in Table 7. Correlations involving Booklets 1A, 1B, and 1 give information about the stability of parameters over a constant context. Correlations involving Booklets 2 to 7 give information about the stability of parameters over a varying context.

Booklets 1A, 1B, 2 and 3, 4 and 5, and 6 and 7 all had item parameters estimated on the basis of about 200 examinees. For the item discriminations the change in context produced a consistent reduction in the correlations. The change in context also decreased the stability of the item difficulties for both the 3-parameter and 1-parameter models. By examining the correlations for Booklets 2, 4, 6 and 3, 5, 7, it can be seen that when data were pooled over different contexts, and sample sizes were therefore tripled, there was an increase in the strength of the linear relationship between first- and second-testing item parameters. However, the correlations for Booklets 2, 4, 6 and 3, 5, 7 (which involved parameter estimates based on sample sizes of about 600) were always lower than the corresponding correlations for Booklet 1 (which involved parameter estimates based on N ≈ 400), and usually lower than the correlations for

Table 6
Correlations Between First-Testing Item Parameter Estimates for 40 X+Y Reading Items (Lower Triangles) and 44 X+Y Mathematics Items (Upper Triangles)

Model, Parameter, and Booklet	Booklets 1A	1B	1	2&3	4&5	6&7	2,4, 6&3, 5,7
3-Parameter Discrimination							
1A		.78	.92	.55	.45	.54	.61
1B	.63		.96	.65	.50	.46	.63
1	.90	.88		.64	.48	.51	.64
2&3	.47	.52	.52		.41	.64	.80
4&5	.39	.28	.40	.31		.47	.77
6&7	.16	.04	.11	.29	.36		.83
2,4,6&3,5,7	.40	.31	.39	.67	.74	.67	
3-Parameter Difficulty							
1A		.94	.98	.82	.80	.75	.84
1B	.87		.99	.87	.85	.75	.88
1	.96	.97		.87	.85	.78	.89
2&3	.77	.72	.76		.91	.72	.95
4&5	.83	.73	.80	.69		.76	.96
6&7	.73	.62	.69	.54	.64		.88
2,4,6&3,5,7	.90	.81	.87	.88	.87	.84	
1-Parameter Difficulty							
1A		.98	.99	.88	.86	.85	.90
1B	.95		.99	.89	.88	.87	.92
1	.99	.99		.89	.87	.87	.91
2&3	.76	.77	.78		.95	.83	.97
4&5	.71	.65	.71	.68		.85	.97
6&7	.65	.63	.65	.55	.68		.93
2,4,6&3,5,7	.81	.81	.82	.84	.88	.88	

Booklets 1A and 1B (N ≃ 200). For Reading the 3-parameter and 1-parameter models had similar stabilities for item difficulties; for Mathematics the 1-parameter model produced slightly more stable difficulties. It is not clear if the difficulty parameters of one of the models was affected more by changes in context than the parameters of the other model.

Effect of Linking

The X and Y item parameters from Booklets 2 to 7 were linked by the use of the anchor items. It is possible that there were inadequacies in the linking procedure that caused the reduced correlations between parameters estimated in these booklets and those in Booklet 1. Therefore, a check on the importance of the linking procedure in affecting item parameter correlations was made. Correlations were computed between the item parameters obtained for Set X using the first-testing booklets and the parameters obtained for Set X using the second-

Table 7
Correlations of First-Testing Item Parameters
with Second-Testing Item Parameters
for 40 X+Y Reading Items and 44 X+Y Mathematics Items

First-Testing Booklets	Discrimination	Difficulty	
		3-Parameter	1-Parameter
Reading			
1A	.76	.95	.92
1B	.72	.84	.89
1	.81	.92	.92
2&3	.63	.76	.82
4&5	.60	.81	.81
6&7	.30	.73	.73
2,4,6&3,5,7	.67	.89	.90
Mathematics			
1A	.76	.92	.97
1B	.78	.96	.99
1	.80	.96	.99
2&3	.67	.91	.92
4&5	.59	.91	.92
6&7	.61	.84	.89
2,4,6&3,5,7	.76	.95	.95

Note. Sample sizes used for estimating the second-testing item parameters were 1,660 for Reading and 1,810 for Mathematics.

testing booklet, Booklet 1; analogous correlations were obtained for Set Y. There was no linking procedure influencing these item parameters. These correlations are contained in Table 8. In the vast majority of cases, the correlations in Table 9 fell within the range of correlations in Table 10 produced by the X items and the Y items correlated separately. These results suggest that the linking procedure did not cause the reduction in item parameter correlations from Booklet 1 to Booklets 2 to 7.

Item Parameter Statistics

The means and standard deviations of the first-testing item parameters are contained in Table 9. Differences between booklets in these means and standard deviations can be the result of context and/or sampling effects. For the item discriminations there were systematic context/sampling effects, particularly for Mathematics. The item difficulties displayed substantial context/sampling effects for Reading, but very small effects for Mathematics. For Reading there were larger mean differences in item difficulties for the 3-parameter model than for the 1-parameter model. Table 10 contains the RMSDs between the item parameters estimated with the various first-testing booklets. For the item difficulties the 1-parameter model produced smaller RMSDs than the 3-parameter model.

Table 8
Correlations of First-Testing X or Y Item Parameters
with Second-Testing X or Y Item Parameters for
20 Reading and 22 Mathematics Items

First-Testing Booklets	Discrimination	Difficulty	
		3-Parameter	1-Parameter
Reading			
X Items			
1A	.70	.95	.92
1B	.75	.79	.89
1	.79	.90	.92
2	.58	.90	.91
4	.60	.84	.85
6	.44	.83	.86
2,4,6	.70	.95	.96
Y Items			
1A	.85	.95	.94
1B	.69	.91	.91
1	.85	.95	.94
3	.66	.78	.86
5	.55	.78	.76
7	.17	.63	.58
3,5,7	.62	.85	.83
Mathematics			
X Items			
1A	.72	.90	.98
1B	.65	.97	.99
1	.67	.96	.99
2	.67	.91	.93
4	.60	.93	.94
6	.67	.90	.94
2,4,6	.81	.96	.96
Y Items			
1A	.80	.96	.98
1B	.86	.95	.99
1	.89	.97	.99
3	.76	.91	.91
5	.60	.91	.90
7	.61	.77	.81
3,5,7	.78	.95	.93

Trait Estimates

Context effects. Using the second-testing data, trait estimates were obtained for the X items (θ_X) and the Y items (θ_Y) for all examinees who answered at least a third of the X and a third of the Y items and who did not have zero or perfect scores. The item parameters on which the traits were based were linked in the first testing, as previously described (see Table 4 for sample

Table 9
Means and Standard Deviations of First-Testing Item Parameters for 40 X+Y Reading Items and 44 X+Y Mathematics Items

			Difficulty			
	Discrimination		3-Parameter		1-Parameter	
Booklets	Mean	SD	Mean	SD	Mean	SD
Reading						
2&3	.98	.42	-.15	.93	-.50	.56
4&5	1.11	.54	.43	.71	-.14	.56
6&7	.94	.46	-.14	1.15	-.56	.83
2,4,6&3,5,7	.97	.36	.03	.68	-.40	.55
Mathematics						
2&3	.83	.37	-.10	.86	-.43	.86
4&5	.97	.50	-.10	.95	-.42	1.00
6&7	1.03	.54	-.09	.88	-.36	.91
2,4,6&3,5,7	.89	.36	-.08	.80	-.39	.87

sizes for item parameterization); θ estimates were based on 20 items for Reading and 22 items for Mathematics. Because the item parameters were linked on the basis of the first-testing data, θ_X and θ_Y theoretically were equated. Table 11 presents the relationships between θ_X and θ_Y. For item parameters estimated with Booklet 1A, the 1-parameter model produced higher correlations between θ_X and θ_Y than the 3-parameter model. However, the 3-parameter model produced closer equatings of means and standard deviations than the 1-parameter model. The RMSD/S_θ between θ_X and θ_Y was greater for the 3-parameter than for the 1-parameter model.

Table 10
Root Mean Squared Differences Between First-Testing Item Parameters for 40 X+Y Reading Items and 44 X+Y Mathematics Items

				Difficulty					
	Discrimination: Booklets			3-Parameter: Booklets			1-Parameter: Booklets		
Booklets	4&5	6&7	2,4, 6&3, 5,7	4&5	6&7	2,4, 6&3, 5,7	4&5	6&7	2,4, 6&3, 5,7
Reading									
2&3	.59	.52	.32	.90	1.01	.49	.58	.71	.32
4&5		.59	.40		1.05	.54		.74	.37
6&7			.35			.71			.47
Mathematics									
2&3	.50	.45	.24	.39	.65	.28	.34	.52	.22
4&5		.55	.33		.63	.30		.54	.25
6&7			.34			.42			.34

The results for Booklet 1A can be compared to those for Booklets 2 and 3 to examine the extent of context effects. The change in context between the parameter estimation and the trait estimation did not systematically affect the correlations between θ_X and θ_Y. The change in context did decrease the closeness of the equating of the means and standard deviations and increase the RMSD/S_θ for Reading but had little effect on Mathematics.

For Booklets 2, 4, 6 and 3, 5, 7 the 1-parameter model produced higher correlations between traits than the 3-parameter model. The means and standard deviations were equated with about equal accuracy for the two models. The difference between the models in correlations was reflected in the lower RMSD/S_θ for the 1-parameter model.

Effects of Unequal Item Difficulties

In order to examine the equating of traits based on items of unequal difficulty, the X and Y items were divided into sets of easier (E) and harder (H) items. This division was based on the proportions of examinees who passed the items in the second testing. The distributions of item difficulties overlapped for the E and H sets, but the mean difficulties for the two sets differed by about as much as is common for adjacent levels of standardized achievement tests. There were 20 items in the E set and 20 items in the H set for Reading and 22 items in each of the two sets for Mathematics. The item parameters for these sets were those obtained in the first testing on the basis of the pooled data for Booklets 2 to 7, and trait estimates were based on second-testing data.

Table 12 presents the relationships between the traits based on the E and H sets. The quality of the equating of the E and H sets can be compared to the equatings of the X and Y sets in Table 11 for Booklets 2, 4, 6 and 3, 5, 7. The correlations for θ_E and θ_H were very similar to the corresponding correlations for θ_X and θ_Y. For Mathematics with the 3-parameter model, the θ_E and θ_H equating was only slightly worse than the corresponding θ_X and θ_Y equating. For Reading for both models and for Mathematics with the 1-parameter model, the θ_E and θ_H equatings of means and standard deviations were noticeably poorer than the θ_X and θ_Y equatings.

For Reading the 1-parameter model produced higher correlations and lower RMSD/S_θ for θ_E and θ_H than the 3-parameter model, whereas the 3-parameter model produced slightly better equatings of means and standard deviations. For Mathematics the 3-parameter model produced closer equatings of means and standard deviations, a slightly lower RMSD/S_θ, and an equal correlation of θ_E and θ_H as compared with the 1-parameter model.

Effect of Trait Level

For all the second-testing trait equatings, the difference between the 1-

Table 11
Relationships Between Trait Estimates Based on X Items (θ_X) and Trait Estimates Based on Y Items (θ_Y) for the Second Testing

First-Testing Booklet	P_X	P_Y	3-Parameter				1-Parameter			
			$r_{\theta_X\theta_Y}$	$\frac{\bar{\theta}_X-\bar{\theta}_Y}{S_\theta}$	$\frac{S_{\theta_X}}{S_{\theta_Y}}$	$\frac{RMSD}{S_\theta}$	$r_{\theta_X\theta_Y}$	$\frac{\bar{\theta}_X-\bar{\theta}_Y}{S_\theta}$	$\frac{S_{\theta_X}}{S_{\theta_Y}}$	$\frac{RMSD}{S_\theta}$
Reading (N=1,525)										
1A			.70	-.02	1.05	.78	.78	-.10	1.06	.67
2&3			.72	.25	.85	.80	.78	.33	1.07	.74
2,4,6&3,5,7	.60	.58	.71	-.03	.91	.77	.78	.03	1.11	.67
Mathematics (N=1,778)										
1A			.75	.08	1.00	.71	.77	.10	.98	.69
2&3			.75	.04	.98	.71	.76	.08	1.02	.70
2,4,6&3,5,7	.58	.60	.75	-.03	1.01	.71	.77	-.02	1.04	.68

Note. P_X is the mean proportion passing the X items, and P_Y is the mean proportion passing the Y items.

$$S_\theta = \sqrt{(S^2_{\theta_X} + S^2_{\theta_Y})\ /2}$$

Table 12
Relationships Between Trait Estimates Based on Easier Items (θ_E) and Trait Estimates Based on Harder Items (θ_H) for Second Testing

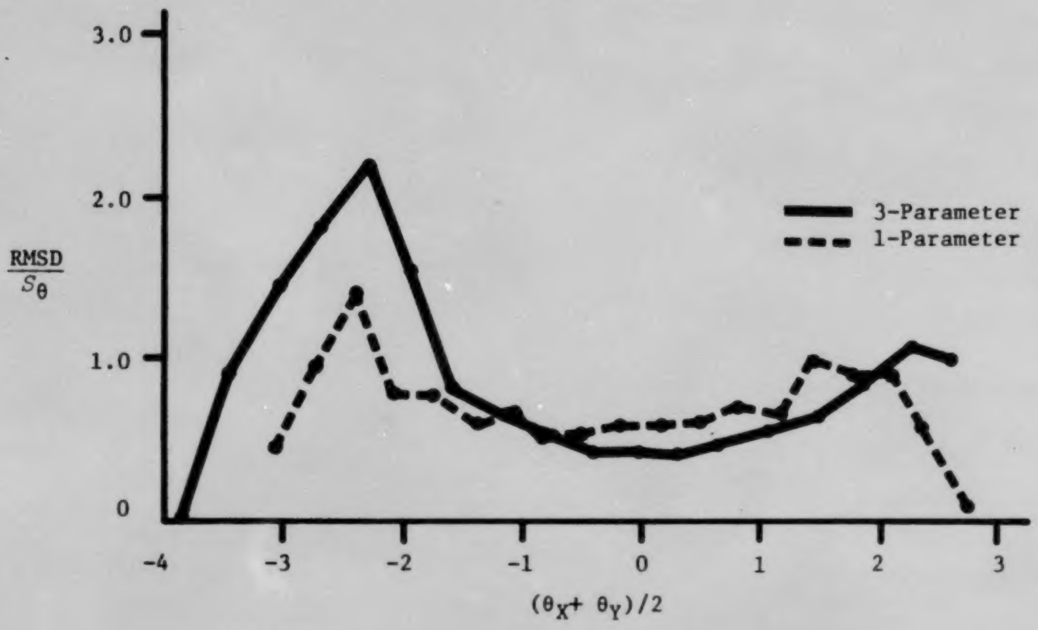
First-Testing Booklet	P_E	P_H	3-Parameter				1-Parameter			
			$r_{\theta_E\theta_H}$	$\frac{\bar{\theta}_E-\bar{\theta}_H}{S_\theta}$	$\frac{S_{\theta_E}}{S_{\theta_H}}$	$\frac{RMSD}{S_\theta}$	$r_{\theta_E\theta_H}$	$\frac{\bar{\theta}_E-\bar{\theta}_H}{S_\theta}$	$\frac{S_{\theta_E}}{S_{\theta_H}}$	$\frac{RMSD}{S_\theta}$
Reading (N=1,525)										
2,4,6&3,5,7	.65	.52	.72	.12	.81	.78	.79	.14	1.25	.69
Mathematics (N=1,778)										
2,4,6&3,5,7	.66	.52	.75	-.05	1.02	.71	.75	-.08	1.12	.72

Note. P_E is the mean proportion passing the Easier items, and P_H is the mean proportion passing the Harder items.

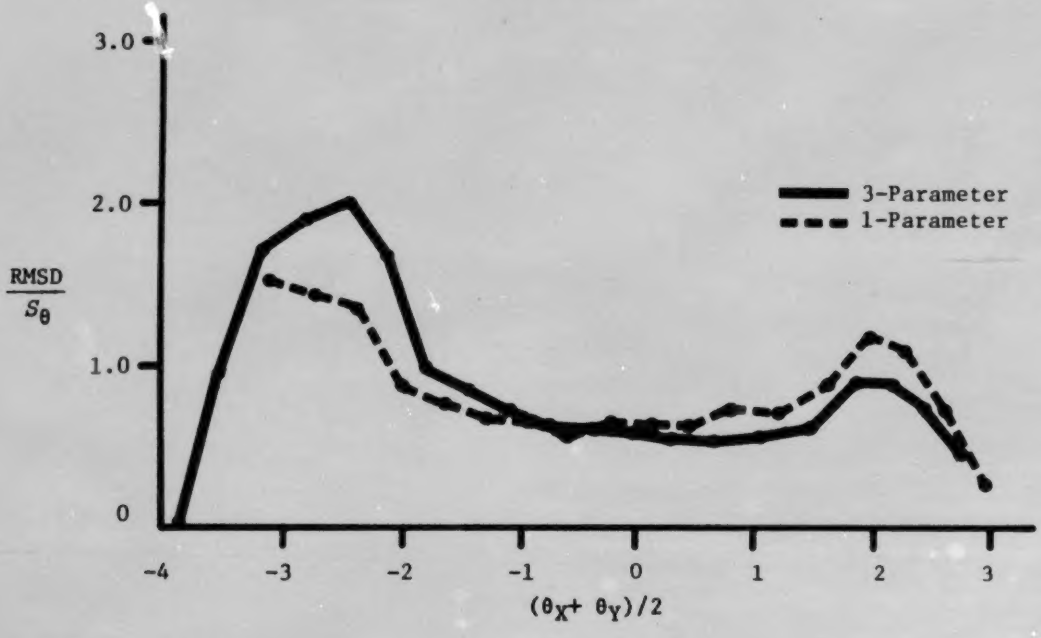
$$S_\theta = \sqrt{(S^2_{\theta_E} + S^2_{\theta_H})\ /2}$$

Figure 1
Second-Testing RMSD/S_θ for θ_X vs. θ_Y, Based on Item
Parameters Estimated Using First-Testing Booklets 2, 4, 6 & 3, 5, 7:

(a) Reading



(b) Mathematics



and 3-parameter models in terms of correlations and RMSD/S_θ was essentially the result of relatively large between-trait differences for low trait values estimated by the 3-parameter model. To display this effect, the average trait estimate for the X items and the Y items, $(\theta_X + \theta_Y)/2$, was found for each examinee. The range of these values was divided into 20 cells. Examinees were sorted into cells on the basis of their mean trait values, and within each cell the RMSD/S_θ between θ_X and θ_Y was found. Figure 1 contain plots of these RMSD/S_θ for traits based on item parameters estimated using Booklets 2, 4, 6 and 3, 5, 7. For low trait values, the RMSD/S_θ was much greater for the 3-parameter model than for the 1-parameter model; these trait values correspond to number-correct scores below those expected by random guessing. The RMSD/S_θ was lower for the 3-parameter model than for the 1-parameter model for cells which included about 70% of the examinees for Reading (Figure 1a) and about 60% of the examinees for Mathematics (Figure 1b).

Observed versus expected proportion passing an item. Cross-validations of the model predictions were made using chi-squares from the examinees' item responses in the second testing; these item responses produced the observed proportions passing the items (O_{ij}). The expected proportions passing the items (E_{ij}) were found using the item parameters estimated in the first testing and the traits estimated in the second testing that were based on the first-testing item parameters and the second-testing item responses. (These trait estimates produced the results in Table 11.) The chi-squares were obtained for the X items and for the Y items. The means (taken over items) of these chi-squares appear in Table 13.

Table 13
Mean Item Chi-squares for the X+Y
Items for the Second Testing

First-Testing Booklets	Model 3-Parameter	 1-Parameter
Reading (N=1,525)		
1A	35	60
2&3	54	63
2,4,6&3,5,7	37	53
Mathematics (N=1,778)		
1A	34	53
2&3	60	72
2,4,6&3,5,7	43	57

Because no item parameters were estimated from the data on which the chi-squares were based, each item chi-square had 10 degrees of freedom for both the 1- and 3-parameter models. Comparing the mean chi-squares for Booklets 1A and 2 and 3, it can be seen that when there was a change in context from the first to the second testing, chi-squares were higher than when the context was constant from the first to the second testing. Pooling over contexts and increasing sample sizes for the item parameter estimates (Booklets 2, 4, 6 and 3, 5, 7) decreased the chi-squares below the level found for Booklets 2 and 3, but usually not below the level for Booklet 1A. The chi-squares for the 3-parameter model were lower than the chi-squares for the 1-parameter model. An examination of the observed and predicted proportions of examinees passing the items revealed that, on the average, the 3-parameter model made more accurate predictions than the 1-parameter model.

Discussion and Conclusion

Item Parameters

It was a consistent finding that X + Y item parameters estimated from the same booklet were more highly correlated than X + Y item parameters estimated from different booklets. There are several factors that could have produced this finding:

1. The inclusion of extra items (Sets A, V, or W) in some of the booklets;
2. The linking of parameters from different booklets by the use of anchor items;
3. Differences in the sample size used for the parameter estimations;
4. Interactions between the ability level of a sample and the parameter estimations;
5. Differences in the number of items scaled together;
6. Systematic differences in the sequence in which items appeared in different booklets; and
7. Unspecified context effects other than sequence.

The evidence for and against the importance of these factors is examined as follows.

Inclusion of extra items (Sets A, V or W) in some of the booklets. Booklets 4 to 7 contained anchor items (Set A items) that were not included in Booklet 1. It is possible that these items altered the trait measured by the booklet. For example, imagine that the Mathematics Set A items were all graph-reading items and that no graph-reading items appeared in Sets X and Y. (This did not occur, but it gives an extreme example of how the Set A items could alter the trait being measured.) To all appearances, the Set A items did not seem to have content systematically different from the X + Y items, but it is possible that they were statistically different from the X + Y items. If the Set A items did alter the trait being measured, then the item parameters could have been affected. If this occurred, the correlations between item parameters estimated in Booklets 4 and 5 and 6 and 7 ($r_{4\&5,6\&7}$) should have been higher than $r_{1A,4\&5}$, $r_{1B,4\&5}$, $r_{1A,6\&7}$, and $r_{1B,6\&7}$. It would also be expected that $r_{4\&5,6\&7}$ would approximately equal $r_{1A,1B}$. An examination of Table 6 does not support these hypothesized relationships among the correlations.

Booklets 2 and 3 contained not only Set A items but also Sets V and W. Thus, Booklets 2 and 3 differed more in content from Booklet 1 than did Booklets 4 and 5 and Booklets 6 and 7. If the inclusion of extraneous items caused item parameters to change, the $r_{1A,2\&3}$ and $r_{1B,2\&3}$ should be lower than $r_{1A,4\&5}$, $r_{1B,4\&5}$, $r_{1A,6\&7}$, and $r_{1B,6\&7}$. However, an examination of Table 6 reveals that the correlations between items in Booklets 2 and 3 and Booklets 1A or 1B tended to be higher than the correlations between Booklets 4 and 5 or 6 and 7 and Booklets 1A or 1B.

Thus, it does not appear that the inclusion of extra items in Booklets 2 to 7 was a major factor in reducing the correlations between item parameters for those booklets and Booklet 1.

The linking of parameters from different booklets by use of anchor items. Results indicate that the linking procedure did not cause the reduction in item parameter correlations from Booklets 1 to Booklets 2 to 7. It should be noted that the procedure used here for linking the items was chosen as the best procedure from among several others: linking by estimating the item parameters in matched samples, linking by using the first principal component of the anchor item difficulties, and linking by using the mean anchor item difficulties and discriminations. The procedure used here produced, in general, the highest correlations among the item parameters and the best equatings of θ_X and θ_Y.

Differences in sample size used for the parameter estimations. For Reading the sample sizes used for obtaining parameter estimates for Booklets 2 and 3, 4 and 5, and 6 and 7 were smaller than those for Booklets 1A and 1B. It is possible that these sample sizes were sufficiently smaller to have caused the item parameters to be noticeably less stable. Because Booklets 2 and 3 had the highest sample sizes among Booklets 2 to 7, it would be expected that $r_{1A,2\&3}$, $r_{1B,2\&3}$, and $r_{1,2\&3}$ would be higher than $r_{1A,4\&5}$, $r_{1B,4\&5}$, $r_{1,4\&5}$, $r_{1A,6\&7}$, $r_{1B,6\&7}$, and $r_{1,6\&7}$. An examination of Table 6 verifies this pattern of correlations. The differences in sample sizes would also suggest that $r_{2\&3,4\&5}$ and $r_{2\&3,6\&7}$ should be higher than $r_{4\&5,6\&7}$; this pattern of correlations does not appear in Table 6. Furthermore, it should be recalled that for Mathematics the sample sizes were as large or slightly larger for Booklets 2 and 3, 4 and 5, and 6 and 7 than for Booklets 1A and 1B; but the reduction in item parameter correlations observed with a change in booklets appeared for Mathematics, as well as for Reading.

Interactions between ability level of a sample and the parameter estimations. It is possible that the samples of examinees obtained for Booklets 2 to 7 were systematically different from the samples obtained for Booklet 1. For example, severe floor or ceiling effects could affect the accuracy and values of item parameter estimates. However, the distributions of abilities for the first-testing booklets appeared quite similar. The mean proportion of items passed for the various first-testing booklets ranged from .57 to .61 for Reading and from .54 to .59 for Mathematics. These results argue against the sample composition having had an important impact on the item parameter estimates.

Differences in the number of items scaled together. There were the same number of items calibrated in Booklets 1, and 2 and 3, but fewer items were calibrated in Booklets 4 and 5 and 6 and 7. Item parameters may have been estimated more stably when more items were calibrated together. To examine this hypothesis, the parameters for Booklets 2 and 3 were recalibrated excluding the V and W items. In this recalibration, Booklets 2 and 3 had the same number of items as Booklets 4 and 5 and 6 and 7. The resulting parameters for Booklets 2 and 3 were correlated with the second-testing Booklet 1 parameters, and these correlations were compared with the corresponding correlations in Table 7. Only two of the six correlations changed: for Reading the 3-parameter item discrimination and difficulty correlations changed from .63 to .61 and from .76 to .77. Thus, it appeared that the number of items being calibrated did not have an important effect on the stability of the item parameter estimates.

Systematic differences in the sequence in which items appeared in different booklets. The sequence in which items appeared within a booklet could have had an influence on item parameter correlations. If sequence is important, it would be expected that item parameters obtained from booklets with similar item sequences would be more similar than item parameters obtained from booklets with dissimilar item sequences. Recall that Table 3 contains rank-order correlations of item sequences between Booklet 1 and Booklets 2 to 7. These correlations would lead to the expectation that for Reading, $r_{1,2}$ would be greater than $r_{1,4}$, which would be greater than $r_{1,6}$; also, $r_{1,5}$ would be greater than $r_{1,3}$, which would be greater than $r_{1,7}$. For Mathematics the sequences of Set X items in Booklets 2, 4, and 6 had similar correlations with Booklet 1, which would lead to the expectation that $r_{1,2}$, $r_{1,4}$, and $r_{1,6}$ would be similar; for the Set Y items, it would be expected that $r_{1,5}$ would be greater than $r_{1,3}$, which would be greater than $r_{1,7}$. The correlations in Table 8 are partially consistent with these expectations. In particular, the booklets with the most similar item sequences tended to have more highly correlated item parameters than the booklets with the least similar item sequences.

These results indicate that the similarity of item arrangements might have an influence on the similarity of item parameters. Part of this influence could be the result of examinee fatigue or impatience to finish the test. If so, items should be relatively less difficult if they appear at the beginning of a booklet than at the end of a booklet. In Table 8 Reading item parameters for Booklet 7 had particularly low correlations with the parameters for Booklet 1. A passage that appeared at the beginning of Booklet 1 and near the end of Booklet 7 was identified. The items for this passage were all relatively more difficult in Booklet 7 than in Booklet 1. These items also had relatively higher discriminating powers in Booklet 7 than in Booklet 1. It did not appear that speededness was an important factor because 93% of the examinees who answered at least a third of the items in Booklet 7 answered the last item in the booklet. A possible explanation is that a significant number of the examinees who answered questions about this passage near the end of Booklet 7 did not take the care that examinees took when the passage was at the beginning of Booklet 1, and that for Booklet 7 items for this passage were important in discriminating between the higher scoring/more careful examinees and the lower scoring/less careful examinees.

Several other analyses similar to the one described in the previous paragraph were conducted. Items appearing at the end of booklets frequently, but not always, were relatively more difficult than the same items appearing at the beginning of another booklet. Results for item discriminations were not as systematic. Thus, it appeared that the location of an item in a booklet could have, but did not have to have, an impact on the item's parameters.

Unspecified context effects other than sequence. Although the location of items in the booklets appeared to be a partial explanation of context effects on item parameters, location did not appear to be a complete explanation. It is possible that other factors related to context could have influenced the parameters. For example, such factors might be specific to the particular content of items. It is not apparent, however, exactly what these factors would be.

Conclusions. After an examination of seven factors that could possibly have influenced the stability of parameter estimates, the conclusion reached is that context effects are not artifacts but can be the result of an item's location in a booklet and, conceivably, other unexplained context effects. It may not be possible to obtain truly context-free item parameters. However, it may be possible to obtain approximately context-free item calibrations by basing the item calibrations on data pooled over administrations of the items in a variety of contexts.

Traits

Systematic differences in item parameter estimates can be important. For example, suppose that the second-testing trait estimates were based on the X + Y items using parameters from the first testing. Because the mean item difficulties varied as a function of the first-testing booklet and sample (see Table 9), the means of the second-testing traits would also vary. Variations in the estimated item discriminations would influence the standard errors of the traits that would be predicted by the 3-parameter model.

Obtaining equated trait estimates is one of the most important tests of the usefulness of the latent trait models. When item parameters were based on data pooled over contexts (Booklets 2,4,6 and 3,5,7), second-testing trait estimates based on items of approximately equal difficulty (θ_X and θ_Y) were fairly well equated. Trait estimates based on items of systematically different difficulty levels (θ_E and θ_H) were less well equated. It is encouraging that well-equated traits were obtainable despite the presence of context effects on the item parameters, but it is apparent that equating errors can be expected to be greater for vertical, than for horizontal, equating.

One potential use of latent trait models with item pools is in basing an examinee's trait estimate on a subset of the items in the pool and predicting whether the examinee would have passed items in the pool he or she did not take. This provides a method of criterion referencing an examinee's trait value. If context effects influence item parameters, such criterion referencing will be inaccurate.

Models

Recall that one of the criteria in the selection of items for this study was fit of the items to the predictions of a 2-parameter logistic model (in which $c_i = 0$ for all items). The V and W items were chosen from the items that had relatively poor fit to that model; and the A, X, and Y items were chosen from the items that had relatively good fit. In theory, an item that fits the 2-parameter model will fit the 3-parameter model but will not necessarily fit the 1-parameter model. This theory implies that the item selection procedure was biased in favor of the 3-parameter model. In practice, however, this bias did not occur. Selection on the basis of fit to the 2-parameter model had the effect of discarding items that had the poorest chi-squares for both the 1- and 3-parameter models.

The item that were retained for Sets A, X, and Y were among those that had the best fit for both the 1- and 3-parameter models; but these items were not those that systematically had the best fit for either one of the models. The mean item chi-squares for the items that were discarded or placed in Sets V and W were 10.5 (Reading) and 12.3 (Mathematics) for the 3-parameter model and 21.5 (Reading) and 24.7 (Mathematics) for the 1-parameter model. The mean item chi-squares for the items chosen for Sets A, X, and Y were 7.1 (Reading) and 7.7 (Mathematics) for the 3-parameter model and 14.6 (Reading) and 13.8 (Mathematics) for the 1-parameter model. It is clear that the selection of the A, X, and Y sets of items had a much greater effect on the mean of the 1-parameter chi-squares than on that of the 3-parameter chi-squares. It is also clear that the items chosen for Sets A, X, and Y fit the 3-parameter model much better than the 1-parameter model. Even if the items had been chosen on the basis of having the best fit with respect to the 1-parameter model, the chosen items would have had a higher mean chi-square for the 1-parameter than for the 3-parameter model.

For small sample sizes ($N \simeq 200$), the 3-parameter model produced less stable item difficulties than the 1-parameter model. For larger sample sizes, the two models' difficulties were essentially equally stable. This result argues for the use of the 1-parameter (Rasch) model for small sample sizes. However, the trait equatings based on item parameters estimated with small sample sizes were frequently so poor that it does not appear prudent to use either model with small sample sizes.

The two models differed in the types of errors they displayed in the equating of traits. The 3-parameter model tended to produce more unsystematic or random error than the 1-parameter model for low trait values (i.e., trait values associated with number-correct scores below those expected by random guessing). The 1-parameter model tended to produce greater systematic errors in trait equatings than the 3-parameter model, as exhibited in the quality of the equating of means and standard deviations.[3]

When the predictions of the latent-trait models were evaluated in a type of cross-validation (see the mean chi-squares in Table 13), the 3-parameter model produced more accurate predictions, on the average, than the 1-parameter model.

---

[3] For an explanation of these results, see Lord (1980), in this volume.

This result argues for the use of the 3-parameter model rather than the 1-parameter Rasch model, particularly for multiple-choice tests with few answer choices.

REFERENCES

Allen, M. J., & Yen, W. M. Introduction to measurement theory. Monterey, CA: Brooks/Cole, 1979.

California Achievement Tests, Forms C and D. Monterey, CA: CTB/McGraw-Hill, 1977.

California Achievement Tests Examiner's Manual, Levels 14-19, Forms C & D. Monterey, CA: CTB/McGraw-Hill, 1977.

Lord, F. M., Symposium and discussion. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis, University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum RM-76-6). Princeton, NJ: Educational Testing Service, 1976.

EFFECTS OF SAMPLE SIZE ON LINEAR EQUATING OF ITEM CHARACTERISTIC CURVE PARAMETERS

MALCOLM J. REE AND HARALD E. JENSEN
AIR FORCE HUMAN RESOURCES LABORATORY

The application of the technology of computer-driven adaptive testing requires the development of large banks of test items. Each bank may contain 250 to 400 items, and all must measure the same ability on the same metric or scale. It is unreasonable and impracticable to assemble a single group of 2,000 subjects for 250 to 400 minutes in order to obtain data on all the items; therefore, a method for linking together subsets of items administered to varying groups must be investigated. Item characteristic curve (ICC) theory offers a unique method of linking subsets of test items due to the invariance property of the ICC parameters. This invariance property rests on the two major theoretical assumptions of latent trait theory: (1) unidimensionality and (2) local independence.

Unidimensionality means that only a single ability is being measured and is assumed to be the property of an item pool, even when assembled into subsets. Local independence means that testees' responses to an item are independent of their responses to another item. More simply stated, this means that an item response is a function of ability and no other factor. In effect, this is a restatement of the unidimensionality assumption. If an item pool is unidimensional, then any shift in score metric that is due to a linear transformation may be corrected or adjusted by application of the proper complementary linear transformation. This is what is meant by the idea that latent trait parameters are invariant to a linear transformation, and it is this theoretical property that allows item pools to be linked and transformed to a common metric.

In previous research efforts, item pools have been linked by the method of linear equating (see Lord, 1977; Ree, 1977; Sympson & Ree, in press) with apparent success. To date there has been little research on the efficacy of these linking procedures and the effects of errors in ICC parameter estimation on their (linearly) transformed values.

ICC Parameters

The three-parameter logistic model of Birnbaum (1968) is the most frequently used for relating item responses to testee ability. The three parameters--a, b, and c--are item discrimination, item difficulty (or location), and probability of chance success (or lower asymptote), respectively.

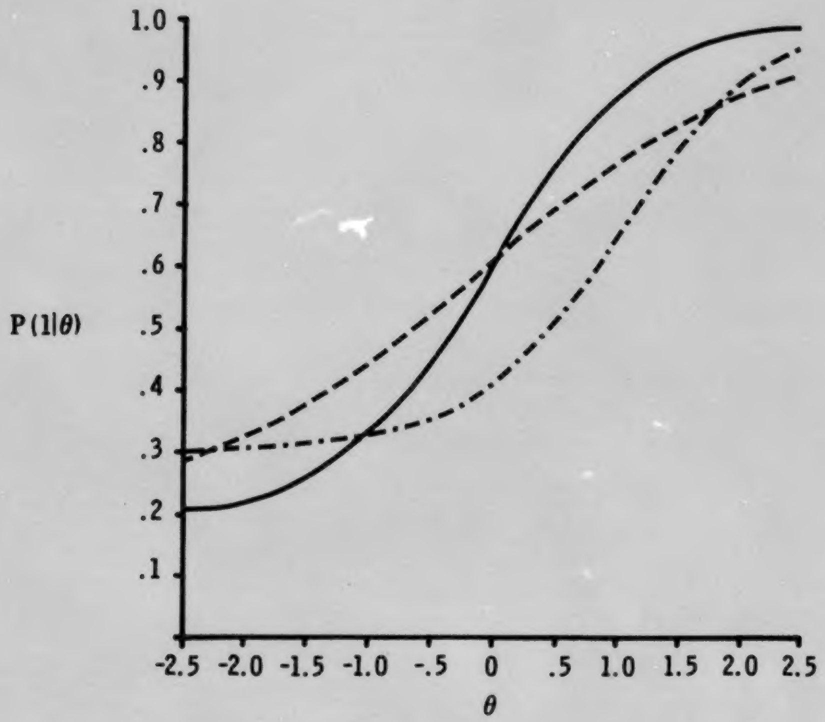
The curve described by these parameters takes the shape of an (cumulative frequency) ogive or an "S," with the upper asymptote approaching a probability of 1.0 and, usually, with a lower asymptote of a probability greater than 0.0. The ogive describes the probability of obtaining a correct answer to an item as a monotonic increasing function of ability.

The item discrimination parameter (a) is a function of the slope of the ICC and generally ranges from .5 to about 2.5. The value of a equal to about 1.0 is typical of many test items, a values below .5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are infrequently found.

The item difficulty parameter (b) describes the point of inflection of the ICC and is usually scaled between -2.5 and +2.5, although the metric is arbitrary. The item guessing parameter (c) is the lower asymptote of the ICC and is generally interpreted as the probability of selecting the correct item option by chance alone. Most test items have c parameters greater than 0.0 and less than or equal to .30.

Figure 1 shows three ICCs. The horizontal axis is scaled in units of ability, θ, and the vertical axis is the probability of answering the item correctly [P(1|θ)]. The solid curved line shows an ICC for an item of average difficulty with acceptable discrimination and the lower asymptote appropriate for a five-item multiple-choice item. The dashed line shows an item of identical diffi-

Figure 1
Item Characteristic Curves



culty, with a *c* value of .28, but a lower *a* value. Note how the slope of the curve is less steep. The third curve, dot-dash line, shows an item with a *c* value of .30, an *a* parameter of 1.0, and a *b* parameter equal to 1.0. As the *b* parameter changes, the location of the inflection point of the curve is displaced along the horizontal axis.

Equation 1 presents the mathematical function describing the curve.

$$P(\theta)_j = c_i + (1 - c_i)\,(1 + e^{(-1.7a_i(\theta - b_i))})^{-1} . \qquad [1]$$

Previous research (Ree, 1978) indicates that the ICC parameters may be estimated with some reasonable degree of accuracy, providing a sufficient sample of examinees with an appropriate distribution of ability (θ) is available.

Linking Paradigms

Two fundamental linking procedures may be defined and are known as the Anchor Items Method (AIM) and the Anchor Subjects Method (ASM). In AIM every subset of items is administered to a different sample of subjects, but embedded into the group of items to be analyzed is a common (or anchor) set of items. During analysis the anchor items are identified, and the following linear transformation is applied to the resultant ICC parameters:

$$b_t = \left(\frac{sb_t}{sb_2}\right) b_2 + \left[\bar{b}_t - \left(\frac{sb_t}{sb_2}\right) \bar{b}_2\right] \qquad [2]$$

where b_t is the item location parameter transformed to the desired scale and sb_t and sb_2 are standard deviations of the desired scale and observed scale, respectively. A similar procedure for the *a* parameter is defined by

$$a_t = a_2 \cdot \frac{sb_2}{sb_t} , \qquad [3]$$

where

a_t is the item discrimination parameter transformed to the desired scale;
a_2 is the observed *a* parameter; and
sb_t and sb_2 are as in Equation 2.

Because the *c* parameter is measured on the probability axis, it does not change and no transformation need be applied.

The ASM requires that the same group of subjects be available to take each subset of items. It is extremely unlikely that the same 2,000 subjects could be assembled to take items over a long period of time, as would be required to place tests on the same metric from year to year. For this reason the ASM method seems less likely to find long-term practical application. Because of its potential for use, the AIM procedure is the subject of the present study.

Method

In order to have a known standard for reference, a simulation study was run using 2 groups of simulated testees, a single set of 20 anchor items, and 2 differing groups of 60 experimental, or nonanchor, items. These two groups of items were assembled into two tests. Both groups of simulees, designated S1 and S2, were specified to have about the same normal distribution of θ. Table 1 shows the mean, standard deviation, and minimum and maximum of θ for Groups S1 and S2. These two groups represent what might be expected if subjects for experimental testing were chosen from a larger pool, such as candidates for military enlistment. Response vectors for these simulees were generated on the two tests.

Table 1
Mean, Standard Deviation, and Minimum and Maximum of θ for Groups S1 and S2

Statistic	Group S1	Group S2
Mean	-.014	.025
SD	.998	1.004
Minimum	-2.600	-2.600
Maximum	2.600	2.600

Generation of Item Responses

In order to generate a vector of item responses for each simulee, the θ values were used in Equation 1 to compute the likelihood of correctly answering each item.

Because Equation 1 yields a number $P(\theta)_j$ such that $0.0 < P(\theta)_j < 1.0$, a number X_j was drawn from a uniform (rectangular) distribution ranging from 0.0 to 1.0 and compared to $P(\theta)_j$. If X_j was larger than $P(\theta)_j$, then an incorrect response was specified for the item; otherwise, a correct response was specified. Thus, a simulee with $P(\theta)_j=.90$ would answer an item correctly 9 in 10 times, and a vector of item responses was developed for each simulee in each data set. These response vectors were then used to investigate the AIM linking procedures.

Table 2 shows the distribution of ICC parameters for the 80 items for Test 1 (T1) and Test 2 (T2), and Table 3 shows the ICC parameters for the 20 anchor items common to both tests.

Simulees from Group S1 were administered only the items in Test 1, and sim-

Table 2
Mean and Standard Deviation of the Generated Item Parameters for Tests 1 and 2

Item Parameter and Statistic	Test 1	Test 2
a		
Mean	1.056	1.045
SD	.279	.239
b		
Mean	.085	-.056
SD	.844	.858
c		
Mean	.188	.202
SD	.054	.047

ulees from Group S2 only the items in Test 2. In order to study the effects of sample size, the ICC parameters were estimated on 4 samples drawn with replacement as follows: 250, 500, 1,000, and 2,000. The ICC parameters were estimated on these 4 sample sizes for both groups. Anchor ICC parameter values from the 4 samples administered Test 1 served as the input values for the anchor item parameters to the second test. This permitted the 4 sizes of the calibration sample (250, 500, 1,000, 2,000) to be varied and to be applied in the 4 samples used to estimate the anchor item ICC parameters.

Results

Table 4 shows the intercorrelations between the known item parameters and the estimated parameters. As past research indicates (Urry, 1976), correlations increased with increasing sample size. The correlations in Test 1 for b and estimates of b started high, at .952, and increased to an exceptionally high .992. Correlations for a and estimates of a began moderately, at .666, and climbed to .869; but the correlations of c and estimated c increased from only .031 to .115. In Test 2 much the same pattern was observed except that the correlation of c and estimated c increased from .164 to .315 as sample size increased.

Because correlations are insensitive to constant differences, as might be found if ICC parameters were either over- or under-estimated by a constant amount, summed absolute deviations of the estimated parameters from the known parameters were computed for each parameter in each sample size. Table 5 presents the summed absolute deviations (or summed errors) for both tests with the four sample sizes. Figure 2 displays this graphically.

There was a large drop in summed error when the a parameter was estimated on progressively larger samples up to and including the difference between 1,000 and 500 simulees. Between 1,000 and 2,000 simulees the difference in summed

Table 3
ICC Item Parameters of the 20
Anchor Items Common to Both Tests

Anchor Item Number	ICC Item Parameter		
	a	*b*	*c*
1	.80	-1.50	.10
2	.80	-1.35	.10
3	1.00	-1.20	.15
4	1.00	-1.05	.15
5	1.10	-.90	.20
6	1.20	-.75	.20
7	1.20	-.60	.22
8	1.20	-.45	.20
9	1.30	-.30	.20
10	1.40	-.15	.20
11	1.40	.15	.22
12	1.30	.30	.25
13	1.20	.45	.20
14	1.20	.60	.22
15	1.10	.75	.22
16	1.00	.90	.20
17	1.00	1.05	.25
18	.80	1.25	.25
19	.80	1.35	.25
20	.80	1.50	.25
Mean	1.06	.00	.20
SD	.21	.95	.04

error was smaller. The relationship between error and sample size for the *b* parameter was more nearly constant. That is, the line on the figure for estimates of *b* is generally straight, which means error tended to be reduced in direct proportion to the number of simulees. The almost flat line for the *c* parameter indicates that virtually no reduction of error occurred with increasing sample size for that parameter. The average absolute deviation for the *c* parameter was almost one-third of the entire range of the parameter, as the *c* parameter is generally estimated between .00 and .30. However, past research (Ree, 1979) indicates that even for low-ability subjects, the effects of errors in the estimation of the *c* parameter are small.

Summed deviations of known ICC parameters from the equated value of the ICC parameters were computed for the *a* and *b* parameters for the 16 combinations of calibration sample size and equating sample size. Table 6 shows the summed absolute deviations and the per item deviation for both parameters for the 16 combinations. The equated *a* parameter showed large summed deviations whenever the sample was limited to 250 simulees, whether in the calibration or the equating sample. The lowest error rates for the *a* parameter occurred when the anchor item values were estimated on 2,000 simulees. The effects of the size of the

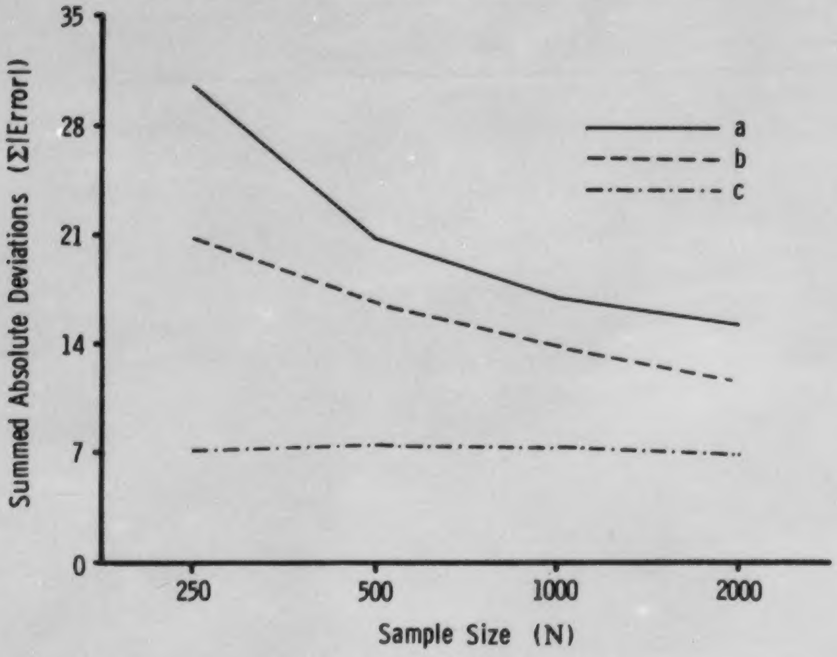
Table 4
Intercorrelations between Known and Estimated ICC Item Parameters for Tests T1 and T2 for Both Groups with Varying Sample Sizes

Item Parameter and Sample Size	Test 1	Test 2
a		
250	.666	.512
500	.671	.725
1000	.831	.813
2000	.869	.886
b		
250	.952	.929
500	.964	.962
1000	.980	.979
2000	.992	.987
c		
250	.031	.164
500	.035	.109
1000	-.012	.331
2000	.115	.315

Table 5
Summed Absolute Deviations ($\Sigma|\text{Error}|$) and Average Absolute Deviations ($|\overline{\text{Error}}|$) for the Three ICC Item Parameters for Tests 1 and 2, for Both Groups with Varying Item Sample Sizes

Item Parameter and	Test 1		Test 2	
Sample Size	$\Sigma|\text{Error}|$	$|\overline{\text{Error}}|$	$\Sigma|\text{Error}|$	$|\overline{\text{Error}}|$
a				
250	30.645	.383	30.529	.382
500	22.809	.285	20.691	.259
1000	15.749	.197	16.891	.211
2000	15.598	.195	15.139	.189
b				
250	23.505	.294	20.847	.261
500	19.860	.248	16.607	.208
1000	17.689	.221	13.805	.173
2000	12.735	.159	11.513	.144
c				
250	7.736	.097	7.235	.090
500	7.360	.092	7.512	.094
1000	6.908	.086	7.318	.092
2000	6.440	.081	6.864	.086

Figure 2
Errors in Estimation of ICC Parameters



calibration sample were not as clear. When 2,000 subjects were used to estimate the anchor item ICC parameters, the magnitude of the error was approximately the same for all calibration sample sizes except 250.

With increasing calibration sample size the error rate increased by some small amount, as indicated by the average (per item) absolute deviation error. This is an unexpected result; an explanation may be found in the relationship between the sets of estimated a parameters. If the estimated a parameters were all estimates of the same value and if the test scale were unidimensional (a basic assumption of the theory), then the estimated a parameters should be linear transformations of one another and should be correlated 1.0, as correlations are invariant to a linear transformation. This was not found to be the case, and Table 7 shows the intercorrelation of the estimated a parameters. Only the correlation between the estimate of a calculated on 1,000 simulees and the estimate of a calculated on 2,000 simulees approached this relationship. This lack of linearity may be due to the assumption of normality and to the rescaling used in the calibration procedure; these may interact in such a way as to produce the anomalous results.

Table 7 also shows the intercorrelations of estimated b parameters. All exceeded .90, and the summed deviations also showed a steady decrease as sample size increased for the b parameter, indicating a virtually linear transformation of estimated b parameters from sample to sample. However, with 500 simulees in the equating sample, a similar anomaly was observed, which may also be due to normal assumptions and to rescaling.

Table 6
Summed Absolute Deviations ($\Sigma|\text{Error}|$) and Average Absolute Deviations ($|\overline{\text{Error}}|$) for Item Parameters a and b for Various Calibration and Equating Sample Sizes

Sample Size		Item Parameter			
		a		b	
Calibration	Equating	$\Sigma\|\text{Error}\|$	$\|\overline{\text{Error}}\|$	$\Sigma\|\text{Error}\|$	$\|\overline{\text{Error}}\|$
250	2000	34.226	.428	23.368	.292
500	2000	15.128	.189	21.934	.274
1000	2000	15.987	.120	16.366	.205
2000	2000	16.596	.207	13.458	.168
250	1000	38.363	.480	25.644	.321
500	1000	17.679	.221	24.341	.304
1000	1000	19.587	.245	19.116	.239
2000	1000	21.032	.263	16.883	.211
250	500	48.611	.608	25.437	.318
500	500	24.558	.307	22.899	.286
1000	500	28.829	.360	18.187	.227
2000	500	31.209	.390	15.833	.198
250	250	44.312	.554	26.201	.328
500	250	21.577	.270	24.416	.305
1000	250	24.439	.312	19.484	.244
2000	250	27.024	.338	17.326	.217

Table 7
Intercorrelations, Means, and Standard Deviations of the Estimated a Parameters (Lower Triangle) and b Parameters (Upper Triangle) for Test 2

	Sample Size				b	
Sample Size	250	500	1000	2000	Mean	SD
250		.952	.940	.935	.056	.856
500	.757		.978	.969	.059	.838
1000	.690	.860		.986	.074	.870
2000	.595	.803	.926		.056	.873
a						
Mean	1.353	1.254	1.235	1.227		
SD	.484	.335	.325	.306		

Discussion

The results of the study present new evidence of the critical interrelationship between item calibration and equating sample sizes and the values of ICC parameters.

Estimating and Equating *a*

For the 16 combinations of calibration sample sizes and equating sample sizes identified in Table 6, the least deviation of estimated a from its known value occurred with an equating sample size of 2,000 and a calibration sample size of 500. As mentioned in the previous section, although the least error between an estimated and known a value was expected with a match of 2,000 equating and 2,000 calibrating sample sizes, error actually increased very slightly with increasing calibration sample sizes beyond 500. This discrepancy apparently resulted from a nonlinear transformation with sample sizes of 250 and 500 but tended toward linearity with sample sizes of 1,000 and 2,000.

During equating procedures a sample size of greater than 500 should be used to ensure an acceptable degree of confidence that the estimation of a does not significantly depart from its "true" value. In the same light, estimation of a suffers considerably using equating sample sizes of less than 500 such that equating samples of 1,000 or 2,000 are highly desirable to minimize error in estimating a.

Estimating and Equating *b*

Table 6 also shows the linear relationship between error and sample size for the b parameter. The b parameter was best estimated with calibration and equating samples of 2,000 each, although a calibration sample size of 1,000 with an equating sample size of 500 can be tolerated without an appreciable increase in error. With all combinations of calibration and equating sample sizes, b was estimated quite well.

Estimating and Equating *c*

The flat line drawn in Figure 2 representing the data from Table 5 shows the estimation of the c parameter to be nearly insensitive to increases in sample size. As sample size increased from 250 to 2,000 subjects, error decreased, but only very slightly. With c defined as the lower asymptote of the ICC and representing the probability of extremely low-ability examinees correctly answering an item, the inability to estimate c with precision could be disturbing. However, it has been pointed out (Lord, 1975) that if $a(\theta - b) < -2$, then the probability of a correct response is c. Therefore, if there are a large number of subjects with ability θ so that $\theta < -(2/a - b)$, c can be accurately estimated. If this requirement is not met, c will be poorly estimated.

Conclusions

A stable and accurate estimate of the a and b parameters requires large numbers of subjects over a broad range of ability. The estimation of c requires large numbers of subjects at very low ability levels. This holds for both equating and calibrating samples; therefore, it is necessary to administer test items, whether to be calibrated or equated, to the largest samples available.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Lord, F. M. *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (ETS RB 75-33). Princeton NJ: Educational Testing Service, 1975.

Lord, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117-138.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Ree, M. J. *Adaptive testing at an AFEES*. Paper presented at the annual meeting of the Military Testing Association, San Antonio, TX, October 1977.

Ree, M. J. *Estimating item characteristic curves* (AFHRL-TR-78-68). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division, November 1978.

Ree, M. J. *The effects of errors in the estimation of item characteristic curve parameters*. Paper presented at the annual meeting of the Military Testing Association, San Diego, CA, October 1979.

Sympson, J. B., & Ree, M. J. *A validity comparison of adaptive testing in a military technical training environment*. Brook Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division, in press.

Urry, V. A five-year quest: Is computerized adaptive testing feasible? In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U.S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington, DC: U.S. Government Printing Office, 1976. (Superintendent of Documents Stock No. 006-00940-9)

DISCUSSION: SESSION 5

GAIL IRONSON
BOWLING GREEN STATE UNIVERSITY

Two key questions seem to arise from the three papers on item linking and equating. Can information be obtained from various combinations of administering subsets of items to samples of people in order (1) to put the items' parameters on a single scale and (2) to put all examinees' trait estimates on a common scale? To the extent that these two questions can be answered affirmatively, different forms of a test can be given to different examinees. For the first question, dealing with the invariance of the parameter estimates, the papers considered sample size, context effects, and the choice of latent trait model. For the second, which deals with equating the traits, there can be a consideration of whether latent trait theory improves equating over classical models (and in what circumstances), a consideration of the problems in vertical versus horizontal equating, and a comparison of the efficacy of the models. These two questions will be discussed interchangeably.

At various points in the studies presented, an anchor test design was used. Two different groups of people each get a basically different set of items, but some of the items are the same. The items commonly given to both groups are called "anchor items."

The question of parameter invariance will be considered first. Theoretically, the item characteristic curve (ICC) parameters estimated using different samples of subjects should be invariant within a linear transformation. Measures of the accuracy of the invariance of parameters are typically the correlation coefficient and some type of error function. In the Ree and Jensen study, an anchor subset of items was embedded in two otherwise different tests administered to two different samples of simulated examinees. Both groups of simulated subjects were specified to have about the same normal distribution of ability. Since it was a simulation study, the true ICC parameters were used. Ree and Jensen looked at both the correlation coefficients and the summed absolute deviations of estimated parameters from the true parameters.

There are two overall trends that have been noted previously in the literature. The first is the relative accuracy in estimating the a, b, and c parameters. The second concerns the sample size, and here there are some consistencies and some inconsistencies. In looking at the correlations, it is found that the correlations for the b parameters run in the mid 90s, even with a small sample size of about 250. The correlations for the a parameters range from the mid 60s to the mid 80s, and it seems that the effect of increasing the sample size is most potent for this parameter. (Unfortunately, the exact function for the a parameter has not yet been determined. Is it a log function, a linear function, or something else?) And the c parameter is not estimated well at any sample size, the correlations ranging from .03 to .31.

With respect to the summed absolute deviations, the error in estimating a is largest, but it seems to decrease more rapidly than for the other parameters. The error for b is less, and it decreases more slowly. And, as mentioned above, even with an increase in sample size, there is no reduction in error for the c parameter. Although the c parameter has the smallest error, the error is about one-third of the range, so that the relative error is quite large.

In Yen's study the correlations between the parameters for various sample sizes can also be examined. With a sample size of 200, different samples took the test in the same context; in other words, it was a sample that was just split into two. These were Yen's 1a and 1b samples. (This can be compared to a sample size of 250 in the study by Ree and Jensen.) The a parameters correlated .63 and .78, which is approximately the same level of correlation observed previously. The b parameters correlated .87 and .94 for two tests--Reading and Math--and the Rasch difficulty parameters correlated higher, .95 and .98. Her study also showed that the change in the context substantially decreased the stability of the parameters: The correlations dropped. This was particularly true for the a parameters. As an example of a change in context, she also noted that the items at the end of the booklet seemed to be relatively more difficult than when they appeared at the beginning of the booklet.

In comparing the 1-parameter and the 3-parameter models, she found that the Rasch model produced slightly more stable difficulties. But it is not clear, according to Yen, whether the parameters of one of the models were affected more by changes in context than the parameters of the other model.

In comparing the joint effect of the sample size and the effect of context, she noted that increasing the sample size increased the correlations between parameters when the context was held constant. However, when pooling over different contexts to increase the sample size, lower correlations were obtained compared to the correlations with smaller samples in the same context. For example, correlations based on N = 600 with items in different contexts were lower than correlations based on an N = 400 in the same context. Thus, as the sample size increased, the correlations increased, but this did not compensate for a change in context. Yen summarized this by stating that the X and Y item parameters estimated from different samples were more highly correlated than X and Y item parameters estimated from different booklets (i.e., different contexts). Finally, she presented an excellent discussion of possible reasons for the invariance, eliminating several reasons.

It is possible that context effects may also be present in the study by Marco, Petersen, and Stewart. They found, for instance, that equating may be different depending on whether the anchor test is internal or external, with more error if the anchor test is external. Though there are many reasons why this could be true, one of them might be the change in context. Since they did not use the same items for the different internal and external tests, that may partially explain the finding. However, context effects could be investigated in that regard.

Ree and Jensen used calibration sizes of 250, 500, 1,000, and 2,000 and equating sample sizes at those four levels; altogether they had 16 combinations of calibration sample size and equating sample size. They looked at the summed deviation of the known ICC parameters from the *equated* values. For the various calibration sample sizes, the equated *b* parameter error decreased most when going from 500 to 1,000. For the equated *a* parameter, the error decreased from calibration sample sizes of 250 to 500 and then increased surprisingly as the sample size went up. Although I do not have the answer to why this happened, I think it is a question that needs to be answered. That same table may be used to look at the equating sample size as well as the calibration size. The equated *a* parameter showed smaller deviations as the sample size increased except when the equating sample size was between 250 and 500, where it increased. The equated *b* parameter had less error with the sample of 500 than with either 250 or 1,000. Thus, there are several anomalies in that same table. The others could be seen by rearranging the table, switching the calibration and equating sample size.

Another observation is that it seems to be more important to keep the calibration sample size above 250 in estimating the *a* parameter. Errors were larger with small calibration samples and large equating samples than with large calibration and small equating samples when sample sizes were between 250 and 500. Finally, in looking at the total combined effect, it seems that if calibrating and equating samples of 250 and 250 are compared to calibrating and equating samples of 2,000 and 2,000, there is roughly a 50% decrease in the error.

The second question was concerned with a comparison of the models. In Yen's study the 1- and 3-parameter models were used, but in some of the subsets of items, one of the criteria for selection was fit to the predictions of a 2-parameter logistic model (in which $c_i = 0$). However, it should be noted that there were certainly items that were selected that did not have zero *c*'s. Also, in two of the subsets of items she allowed items that did not fit well. For the items selected on the basis of fit to the 2-parameter model, she noted that this interestingly had the effect of discarding items that had the worst chi-squares for both the 1- and 3-parameter models. The items retained (Sets A, X and Y), however, fit the 3-parameter model better than the 1-parameter model.

For small sample sizes, an N of 200, the Rasch model generally had slightly more stable difficulty estimates than the 3-parameter model. For large samples the difficulties were stable for both models. However, trait equatings based on item parameters estimated with small sample sizes were so poor that she recommended neither model be used.

The two models differed in terms of the types of errors in equating of the traits. The 3-parameter model tended to produce more unsystematic or random error than the Rasch model for low trait values. The Rasch model had greater systematic errors than the 3-parameter model. The Rasch model also had higher correlations between traits and lower root mean square differences, but the 3-parameter model generally had closer equatings of means and standard deviations. Yen noted that this seemed to be essentially the result of relatively large between-trait differences for low trait values estimated by the 3-parameter model. Finally, the 3-parameter model made better predictions in cross-validation.

In the ETS study the 1-parameter model seemed to be superior to the 3-parameter model ICC in equating a test to itself. However, Marco, Peterson, and Stewart noted that there may be a natural bias operating here because the a's and c's were fixed. The 3-parameter model was the best equating model when total tests of unequal difficulty were equated through a medium difficulty anchor test with dissimilar samples.

The question might be asked, which method seems to hold more promise for some of the problems that are investigated in measurement, for instance, horizontal versus vertical equating? In an ideal horizontal equating study the test forms would be at comparable difficulty levels; there might be minor unintended differences in ability level of the samples; and the anchor test would be roughly parallel to the whole test. Under these conditions the conventional methods seem to work. In fact, some, but not all, of these conditions are necessary. For example, to generalize from Petersen's study, when a test is equated to a test like itself through a parallel anchor test, then the linear model yields good results even if different samples are used. If samples of different ability are taken and there are test forms at comparable difficulties, and if the anchor test is different in difficulty from the total test, then the ICC methods would be best--the 1-parameter slightly better than the 3-parameter model.

In the typical situations in vertical equating there would be two test forms that would be different in difficulty, and groups of examinees who would normally differ in ability level. This is, of course, a more difficult problem. The results from Yen's paper suggest that equating errors would be greater for vertical than for horizontal equating because the trait estimates based on items of systematically different difficulty were less well equated. However, she had the same ability level for that particular result, whereas in some vertical equating situations there would be different ability levels.

The results from the Marco, Petersen, and Stewart study suggest that the 1-parameter model would not handle vertical equating very well. The study found that when total tests differ in difficulty, the 1-parameter model gave unacceptable results in many instances. This is also consistent with the findings of Slinde and Linn (1977). It seems that the 3-parameter model holds more promise for vertical equating.

In summarizing some of the evidence on trait equating from Yen, it should be noted, as she pointed out, that obtaining equated trait estimates is one of the most important tests of the usefulness of latent trait models. The first finding is that trait estimates based on items of approximately equal difficulty were fairly well equated. As previously mentioned, the Rasch model had higher correlations between trait estimates and lower root mean square differences, though the 3-parameter model was better for equating the means and standard deviations. The second conclusion that Yen came to was that trait estimates based on items of systematically different difficulty (i.e., an easy versus a hard test) were less well equated. Third, she found that traits can be equated well despite the presence of context effects on item parameters if there are large samples.

The ETS study presented by Marco, Petersen, and Stewart examined the adequacy of five types of score-equating models when certain sample and test characteristics were systematically varied. The equating models were linear models, equipercentile models, and ICC models. The samples were either random or dissimilar. They looked at several variations on the test characteristics, paying particular attetntion to the relationship of the anchoring items to the whole test. The study was basically divided into two parts. The first part was equating a test to itself. The second part was equating a test to a different test. For the first part Marco, Petersen, and Stewart looked at anchor tests that were either internal or external and at anchor tests that were either more difficult or easier than the whole test. In the second part--equating a test to another test through an internal anchor test--they examined the effects of the difficulty of two total tests and the similarity of the equating samples. They had two different methods of obtaining the criterion scores. The extent to which that might have influenced the results is not entirely clear; however, they are fully cognizant of that problem and did say that the findings were tentative.

The results of the investigation are displayed clearly in a series of figures in the paper. In equating a test to itself, they found that for medium difficulty anchor tests of similar content to the total test, the best linear model had the smallest total error, followed by the 1-parameter and 3-parameter models. They also found that if the anchor test is almost parallel to the total tests, the difference between samples is not so important, i.e., whether it is a similar or dissimilar sample. Another finding was that there was less error, in general, with internal anchor tests than with external anchor tests. The second part of the exploration of equating a test to itself examined easy and difficult anchor tests. The effect of similarity of the samples is potent; having similar samples is more important if the anchor test difficulty is off-center, compared to the total test difficulty. The second point is that linear models are still best if the samples are similar; however, ICC models are superior when the anchor test is different in difficulty from the total test and the samples differ in ability.

For equating a test to a different test (of different difficulty) using an internal anchor of similar content and medium difficulty, they found that the smallest error was present for the 3-parameter model; that was followed by the equipercentile method. There was little effect of the sample.

Several questions have been raised by these studies. First is the question why the increase in calibration sizes changed the accuracy of the a parameter during equating in the Ree and Jensen study. Could this possibly have something to do with the Urry program? I really do not know.

In addition to asking how stable the parameter estimates are and how well traits can be estimated on a common scale, we ought to start looking at the characteristics of those items whose parameters are not stably estimated and where on the scale things are not working properly and why. The c parameter seems to be particularly difficult to estimate; we might do better if we "stuffed the ends." For instance, it might be found that the c parameters are not stably estimated when there are not enough low-ability examinees. It might also be found that if a rectangular distribution of ability is used, more stable estimates of the *a* parameters are obtained and even more stable estimates of the *b* parameters might be obtained for those items that are off-center. It might be asked where on the scale things are not working properly for equating of traits, too. For example, Yen found that the 3-parameter model tended to produce more unsystematic or random error than the Rasch model for low trait values--which is a step in the right direction.

The next question is one raised by Marco, Petersen, and Stewart: To what degree are their results influenced by the criterion equating procedure? They also noted that in some of the cases the rank order remained the same, which would, of course, give a little more confidence in terms of the generalizability of the results.

We have looked at equating under various conditions, for example, the characteristics of the test, the anchor items and their relation to the test, and the sample characteristics. Of course, we could continue getting every possible combination and permutation of these, and this would at least keep us busy until the next adaptive testing conference. One combination that Marco, Petersen, and Stewart mentioned was the effect of internal versus external anchors when equating tests of different difficulty. We might also look at the length of the anchor test and see how that affects equating.

What conclusions are to be drawn? The first conclusion is that under optimum conditions, everything works well. Of course, it helps to have a few thousand people at one's disposal. Second is a finding that has been repeated over and over again. The *b* parameter is estimated the best, then the *a* parameter, and the *c* parameter is estimated rather poorly. With small samples it was found that the Rasch difficulty parameter was more stably estimated by the Rasch model than by the 3-parameter model. A third conclusion is that the changes in context have substantial effects on item parameters, but if there is a large sample size for estimating parameters, trait equatings usually are good.

Fourth, as we already know, vertical equating seems to be more of a problem than horizontal equating. Fifth, using various combinations of the whole test characteristics, the anchor test characteristics, and the samples, the conditions under which the various equating procedures work best have been described; there is now a source to consult to find out the best procedure under given circumstances.

As I mentioned before, the papers in this session shed light on two major questions of interest: To what extent are the parameters invariant? and To what extent can traits be estimated by different items and be put on a common scale? Obviously, we still have a long way to go, but I think we are moving in the right direction.

REFERENCES

Slinde, J. A. & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, 1977, *15*, 23-35.

SESSION 6:

LATENT TRAIT MODELS WHICH INCORPORATE RESPONSE TIME AS WELL AS RESPONSE APPROPRIATENESS

A MODEL FOR INCORPORATING RESPONSE-TIME DATA IN SCORING ACHIEVEMENT TESTS

KIKUMI TATSUOKA AND
MAURICE TATSUOKA
UNIVERSITY OF ILLINOIS

LATENT TRAIT SCORING OF TIMED ABILITY TESTS

DAVID THISSEN
UNIVERSITY OF KANSAS

DISCUSSION

JOHN B. CARROLL
UNIVERSITY OF NORTH
CAROLINA AT CHAPEL HILL

A MODEL FOR INCORPORATING RESPONSE-TIME DATA IN SCORING ACHIEVEMENT TESTS

KIKUMI TATSUOKA AND MAURICE TATSUOKA
UNIVERSITY OF ILLINOIS

The study by Tatsuoka and Birenbaum (1979) raised an important issue with respect to adaptive diagnostic testing and computer-managed routing by which each examinee is sent to his/her level of instruction: that it is necessary to consider an alternative scoring procedure in which individual differences in information-processing skills are taken into account along with individual ability or achievement levels.

In Tatsuoka and Birenbaum's study a computerized diagnostic adaptive test for a series of pre-algebra signed number lessons was given to eighth graders at a junior high school, and a computer-managed routing system sent each examinee to the instructional unit corresponding to the level of skill that he/she reached in the initial test. The adaptive test for signed numbers consisted of 12 groups of items representing 12 different skills. The instructional units of computerized lessons teaching the same 12 skills were rearranged into the same order as the skills in the adaptive test, so that if an examinee stopped at the 7th skill level, he/she was sent to the 7th level of the lessons. After the student went through the 7th to 12th instructional units, a 52-item conventional computerized posttest was administered.

Factor analysis revealed that the test scores of the posttest did not satisfy the assumption of local independence, i.e., unidimensionality. A further close investigation was performed by a cluster analysis on the 92 examinees' response patterns on the basis of Euclidean distances between pairs of response vectors. The result of this analysis led to finding a group of students whose response patterns were significantly different from others. Their scores on the items prior to the stopping level of the initial diagnostic test were higher than most scores of other students, but their scores on the subsequent items were as low as the poorest students' scores. It was confirmed with their teachers that most of them were actually "A" students. It was also confirmed that the members of this group were taught signed number addition operations by a teaching method different from that of the subsequent instructional units, which teach subtraction operations. The procedures of information processing associated with these two instructional methods of performing arithmetic upon signed numbers are greatly different. The traditional scoring procedure of latent trait theory would not be capable of detecting these discrepancies associated with different information processes for arriving at the answers to a given item.

A study by Tatsuoka and Tatsuoka (1978) indicated one useful approach toward the goal just mentioned. It showed that under certain general conditions, item response time scores very closely follow Weibull distributions—a 3-parameter family extensively used in system reliability theory (see, e.g., Mann, Schafer, & Singpurwalla, 1974). The most interesting of the three parameters is the shape parameter, whose magnitude determines the nature of the conditional response rate, that is, the conditional probability that an examinee who has not responded to an item up to time $\underline{t}$ will respond to it within an infinitesimally short time interval thereafter. A brief note on the mathematical and conceptual backgrounds of the Weibull distribution, introduced in the study of Tatsuoka and Tatsuoka (1978), will be described in the following section.

As a follow-up to the Tatsuoka and Birenbaum (1979) study, Weibull distributions were fitted to every item in the posttest. The Weibull fit of almost all items--14 items on addition that were taught prior to the students' exposure to the PLATO lessons--was quite poor when the fitting was done for the total sample. However, the separate fits in two groups, which had earlier been identified as having distinctly different instructional backgrounds, were very good for all 14 items (see Appendix Tables A-1, A-2, and A-3). Further, it was found that the value of the shape parameter $\underline{c}$ differed considerably in the two groups for each item, being higher in one group for some items and lower for others. That is, there was a Task × Instructional Method interaction effect on the shape parameter $\underline{c}$.

The foregoing suggests that the Weibull shape parameter can assist in the identification of items that are sensitive to particular information-processing skills. After identifying and constructing such discriminating items, it was anticipated that an index known as person conditional response rate, to be developed below, could be used for postdicting the instructional background of students and routing them accordingly.

Rationale of Weibull Distributions

Measuring the time needed to achieve a given goal (that is, response time) is easy in computer-managed testing; but since it is impossible to collect accurate response time data in paper-and-pencil testing, it has not been utilized thus far in the realm of practical application of psychometrics. Tatsuoka and Tatsuoka (1978) have studied the statistical aspects of response time distributions and their characteristics as associated with test items.

There are a number of theoretical distributions by which the response time data may seem to be fitted well, so it is necessary to follow some guidelines as to what sort of distribution might be appropriate to represent a set of response times for a given item. Rasch (1960) used the 2-parameter gamma distribution as a model for the time taken to read a passage of N words and the Poisson process as a guide to his model. The occurrence of a response is a random event, and all the random events were assumed to be of the same kind. Rasch was interested in their total number.

Application to Ability Testing

Sato (1975) and others introduced the Weibull distribution, which has been used extensively in the context of system reliability theory. Reliability theory is the study of the probability of failure, within a given time span, of a mechanical or electronic system as a function of the probabilities of failure of individual components of the system. The justification for utilizing a distribution from such an alien field is that the test item is identified with the system whose longevity is being assessed. The student's attacks on the item correspond to the shocks or wear and tear to which the system is subjected, and the eventual solution of the item is the failure of the system. It is plausible to imagine the student to be intent on cracking the system by answering the item correctly. The time he/she takes in doing so, the response time, corresponds to the "survival time" of the system. This rationale for the applicability of Weibull distributions for item response time does not lead to a derivation of the distribution or the density function. Mann et al. (1974) and others have said that the distribution was empirically discovered, rather than deductively derived. Later, a logical basis was postulated as an ex post facto rationalization, and it added greatly to the credibility of the distribution in the theory of system reliability. This is the concept of hazard rates, which is essentially the conditional probability that a system that has survived through time t will fail during an infinitesimal time interval immediately after that.

Conditional Response Rate

A similar concept, conditional response rate (CRR), was introduced in this study as a logical basis for use of the Weibull distribution. Suppose $f(t)$ is the probability density that a person randomly selected from the population will respond to a given item during the interval $[t, t + dt]$. Then, the proportion of individuals who will have responded to the item by time t is the probability distribution function $F(t) = \int_{t_0}^{t} f(u)\,du$. The proportion of individuals who have not responded to the item by time t is $1 - F(t)$. Consequently, the conditional probability density that a person will respond to the item during the interval $[t, t + dt]$, given that he or she has not responded to the item up to time t is given by $f(t)/[1 - F(t)]$.

By assuming CRR as a function of time t to be monotonically increasing, or decreasing, as a power function of t, the Weibull distribution and density functions can be expressed as follows:

$$F(t) = \begin{cases} 1 - \exp\left[-\left(\dfrac{t-t_0}{u}\right)^c\right] \\ 0 \end{cases} \qquad [1]$$

and

$$f(t) = \begin{cases} \dfrac{c}{u}\left(\dfrac{t-t_0}{u}\right)^{c-1} \exp\left[-\left(\dfrac{t-t_0}{u}\right)^c\right], \\ 0 \end{cases} \qquad [2]$$

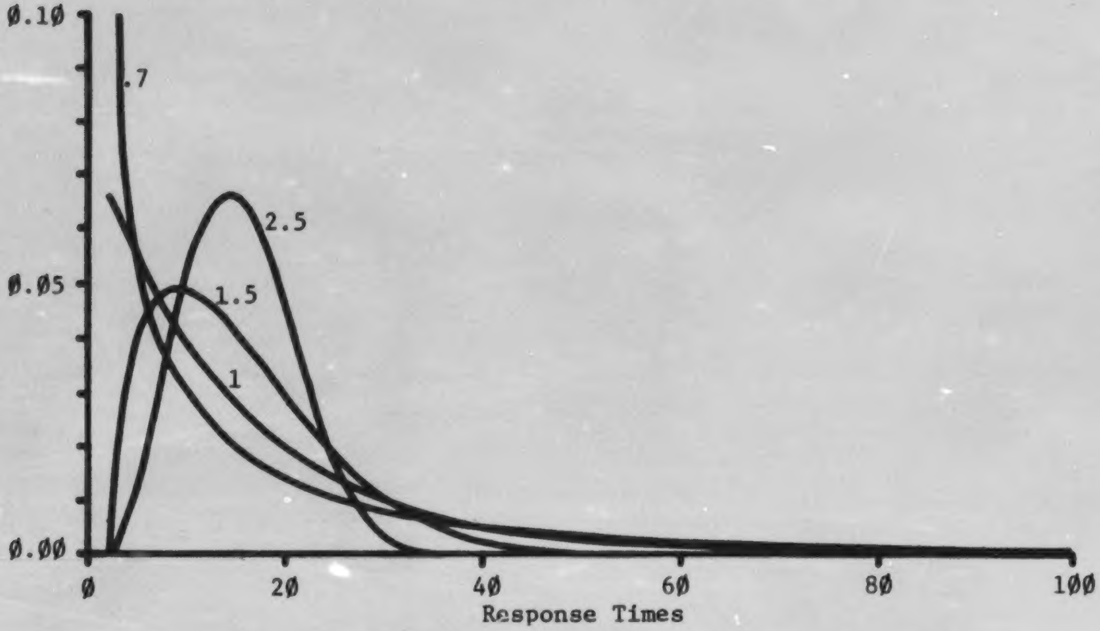
where

$c (>0)$ is the shape parameter,
$u (>0)$ is the scale parameter, and
$t_0 (>0)$ is the location parameter.

Figure 1 shows several Weibull distributions.

Figure 1
Weibull Density Functions with $t_0 = 2$
$\mu_0 = 15$, and Four Values of c

$$f(t) = (c/\mu_0^c)\ (t-t_0)^{c-1} \times \exp[-((t-t_0)/\mu_0)^c]$$

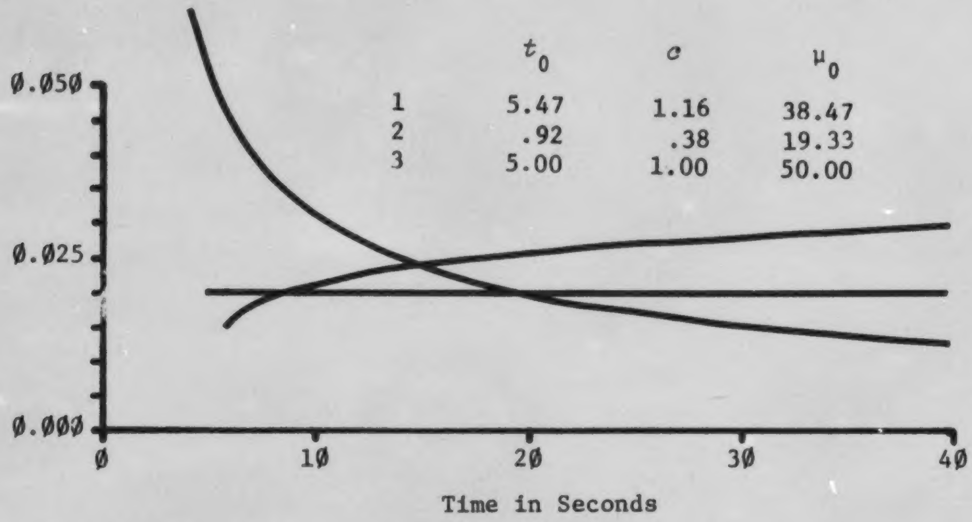


If $c = 1$, then $f(t)$ is a negative exponential density function. If c is less than 1, then $f(t)$ is a monotonically decreasing function. The Weibull density function is symmetric when c is about 3.6. Figure 2 is a CRR function obtained from live data. CRR 1 in Figure 2 is the CRR when c was larger than 1; the decreasing dot graph (CRR 2) was obtained from the distribution when c was less than 1. When $c = 1$, CRR becomes a straight line (CRR 3) that is parallel to the time axis.

Goodness-of-Fit-Tests

Figures 3 and 4 show the displays of goodness-of-fit tests with the normal and Weibull distributions. The step function represents the cumulative distribution of a set of response times to a matrix multiplication. The continuous line stands for the estimated theoretical distribution function. The Weibull distribution fits the data better than does the normal distribution. About 700

Figure 2
Three Types of Conditional Response Rate Function
$F'(t)/[1 - F(t)]$, $F(t)$ = Weibull Distribution Function



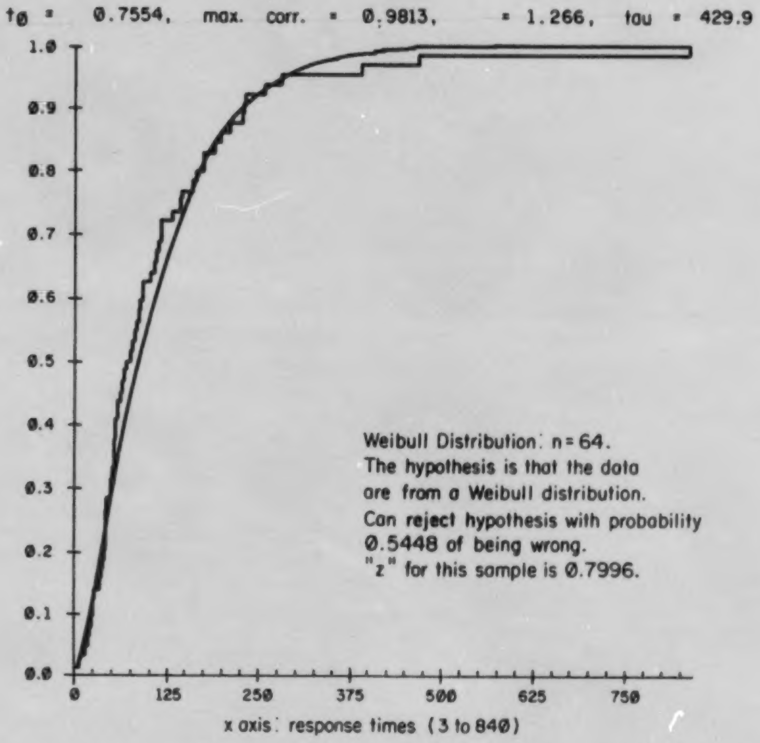
cases of the goodness-of-fit test were carried out, and most data fitted either the Weibull or 3-parameter gamma distributions.

Theoretical distributions were fitted to the observed response time distribution of each item in two ways: (1) for the subgroup of students who answered the item correctly (OK subgroup) and (2) for the subgroup of students who answered the item incorrectly (NO subgroup). The OK subgroup and the NO subgroup had considerably different estimated Weibull parameters, but both showed very good fits for most items. Figure 5 shows the estimated Weibull distributions of the OK subgroup and the NO subgroup for an item in the pretest that required matrix multiplication.

The Weibull parameter $\underline{c}$ of the OK subgroup in a 48-item matrix algebra pretest correlated .32 with the numbers of options in the item and .41 with the difficulty indices. The items with more choice options tended to have large $\underline{c}$ values. If the interpretation that the item $\underline{c}$ value reflects the degree of engagement students show when the item is correct (Tatsuoka & Tatsuoka, 1978), it may be concluded that within the range represented, the larger the number of options, the greater the engagement students feel. This seems reasonable, since items with more options present more of a cognitive task and, hence, probably induce greater involvement on the part of the students. About 10 items in the test asking mathematical properties of orthogonal transformations, eigenvalues, and eigenvectors, were very difficult for many students in the course. These items tended to have the smaller Weibull shape parameters $\underline{c}$ in both OK and NO subgroups. A similar observation was obtained from the 64-item signed number pretest.

The 3-parameter gamma distributions fit well the items that repeatedly re-

Figure 3
Goodness-of-Fit Test for the Time Data and Weibull Distribution Function for Question 17



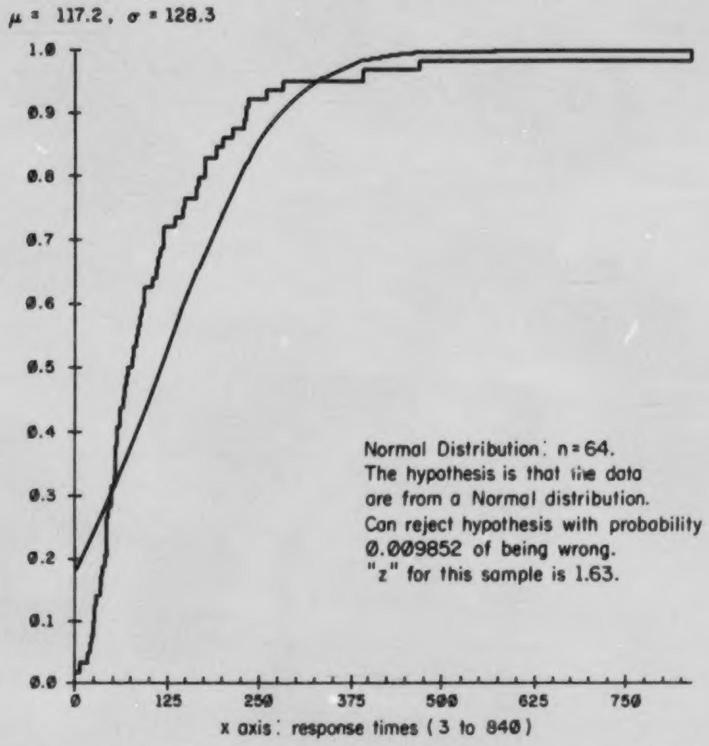
quired a simple mechanical task, whereas the Weibull distributions fit well the items that required a higher cognitive task to respond to. Since the CRR of the gamma distributions is always nondecreasing, that is, either monotonically increasing or parallel to the time axis (see Appendix B), the interpretation of the Weibull shape parameter (see Figure 2) provides wider applicability than the gamma shape parameter does. Moreover, the parameter estimation routine by maximum likelihood usually failed to give convergent estimated gamma parameters when items had decreasing CRRs.

Latent Response Time Model

Latent Response Time Variable and Item Response Time Characteristic Curve

As a first step toward developing the person conditional response rate (PCRR), the existence of a latent response time variable, analogous to the ability variable θ in latent trait theory, is postulated. Thus, given a set of $\underline{n}$ items, the performance on which is affected by θ, it is assumed that there also exists a variable affecting the time taken by an examinee to answer each of these items. There will be no attempt to give any precise psychological meaning to this construct beyond saying that it may be regarded as a pervasive trait of individuals to be slow or quick in solving items of a certain domain.

Figure 4
Goodness-of-Fit Test for the Response Time Data and Normal Distribution for Question 17



Normal Distribution: n = 64.
The hypothesis is that the data are from a Normal distribution.
Can reject hypothesis with probability 0.009852 of being wrong.
"z" for this sample is 1.63.

The plausibility of this postulation is suggested by the following empirical findings. In the Tatsuoka and Tatsuoka (1978) study, the performance scores on a 48-item matrix algebra test were found to have a strong tendency toward unidimensionality. At the same time, the response times for these items showed a suggestion of unidimensionality by the scree test. On the other hand, the posttest for the signed number lessons mentioned earlier showed no semblance of unidimensionality in the total sample. However, when one instructional background group identified by cluster analysis (hereafter called Group 2) was removed, both performance scores and response times came somewhat closer to being unidimensional in the remaining sample.

On the strength of these observations and of the fact, mentioned earlier, that the Weibull distribution fits the response time data for most items, a model for item response time is developed in the following manner, roughly paralleling latent trait theory.

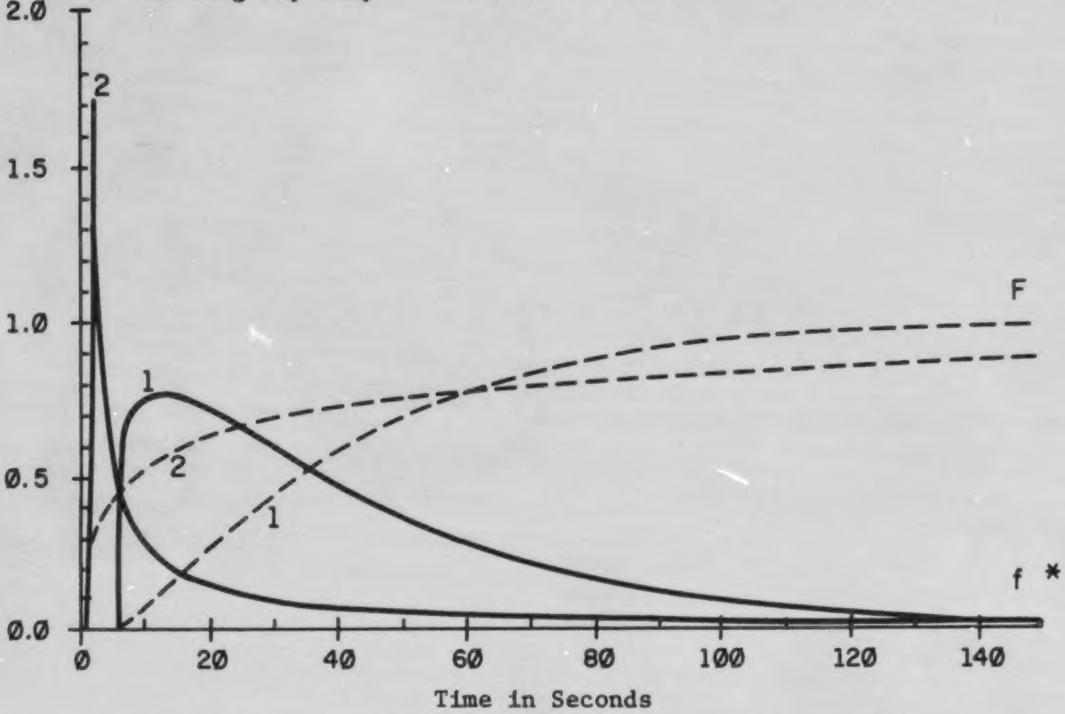
Let

$$d_{ig} = t_{ig} - \bar{t}_{i.} \qquad [3]$$

be the deviation of individual i's response time t_{ig} for item g from his or her

Figure 5
Weibull Distribution and Density Function for Question 16
(Vertical Scale for $f(t)$ is Magnified by a Factor of μ_0)

	c	t_0	μ_0	mean
1 = ok subgroup only	1.161	5.4740	38.47	36.51
2 = no subgroup only	0.3764	0.9206	19.33	76.68



mean response time over the given set of $\underline{n}$ items. Now τ_i may be conceptualized as the expectation $\underset{g}{E}(\underline{t}_{ig})$ of that person's response times over an infinite number of items of the same type as those in the set of $\underline{n}$. Then, $\tau_i + \underline{d}_{ig}$ is approximately equal to $\underline{t}_{ig}$, so if $\underline{i}$ varies across the population, it is reasonable to assume that $\tau_i + \underline{d}_{ig}$ follows a Weibull distribution just as $\underline{t}_{ig}$ does. Therefore,

$$F_{d_g}(\tau) = 1 - \exp\left[-\left(\frac{\tau + d_g}{u_g}\right)^{c_g}\right] \qquad [4]$$

is defined as the response time characteristic function (RTCF) for item $\underline{g}$, where $\underline{t}_0 = 0$ in the general expression Equation 2 for the Weibull distribution function to simplify the task of parameter estimation. This is interpreted to represent the probability that a person whose latent response time is τ will arrive at the answer to item $\underline{g}$ at or prior to time $\tau + \underline{d}_g$.

For estimating the two item parameters $\underline{c}_g$ and $\underline{u}_g$, as well as the person parameter τ_i, the density function corresponding to Equation 4 is written in accordance with Equation 1 (with $\underline{t}_0$ set equal to 0), for each person $\underline{i}$, and the product over all items and those individuals who got the item correct is formed to obtain the likelihood function. That is,

$$L = \prod_{g=1}^{n} \prod_{i=1}^{Ng} f(\tau_i + d_{ig}) = \prod_{g=1}^{n} \prod_{i=1}^{Ng} \exp\left[-\left(\frac{\tau_i + d_{ig}}{u_g}\right)^{c_g}\right] \qquad [5]$$

$$\frac{c_g}{u_g}\left(\frac{\tau_i + d_{ig}}{u_g}\right)^{c_g - 1}$$

where N_g is the number of subjects in the OK subgroup for item $\underline{g}$.

Before going to the next step of developing the PCRR function $G(\tau)$, note how τ itself, once estimated, can help in the task of postdicting a student's instructional background. Suppose there are two items that differentiate between two prior instructions--A and B--by actually showing a reversal in the magnitude order of mean times required for their solution by examinees who were previously taught by these two methods. Table 1 shows the mean response time (also with the estimates of Weibull parameter $\underline{c}$ and CRR) of 14 items described earlier for the two groups, the prior instructional methods of which were A and B, respectively.

Table 1
Means of Response Time, Observed Conditional Response Rate at Mean, and Weibull Shape Parameters of Addition Problems of 64-Item Signed Number Test

Item	Mean Response Time		CRR at Mean		Weibull Shape Parameter	
	Others	Group 2	Others	Group 2	Others	Group 2
3	9.84	13.14	0.13	0.07	1.45	.80
4	7.61	5.13	0.13	0.20	0.99	1.01
14	14.99	18.48	0.10	0.08	1.78	1.68
17	6.39	8.60	0.19	0.11	1.35	0.86
18	8.35	9.00	0.17	0.10	1.68	0.90
19	7.16	8.44	0.18	0.10	1.47	0.75
28	11.78	11.89	0.10	0.13	1.25	1.96
31	8.70	10.85	0.12	0.09	1.00	0.91
32	9.65	5.43	0.09	0.23	0.84	1.43
33	4.62	7.22	0.28	0.17	1.49	1.46
42	14.28	14.85	0.10	0.07	1.76	1.16
46	10.08	9.83	0.10	0.10	0.93	1.03
47	6.22	11.55	0.17	0.07	1.05	0.63
56	10.56	9.69	0.14	0.14	1.87	1.64

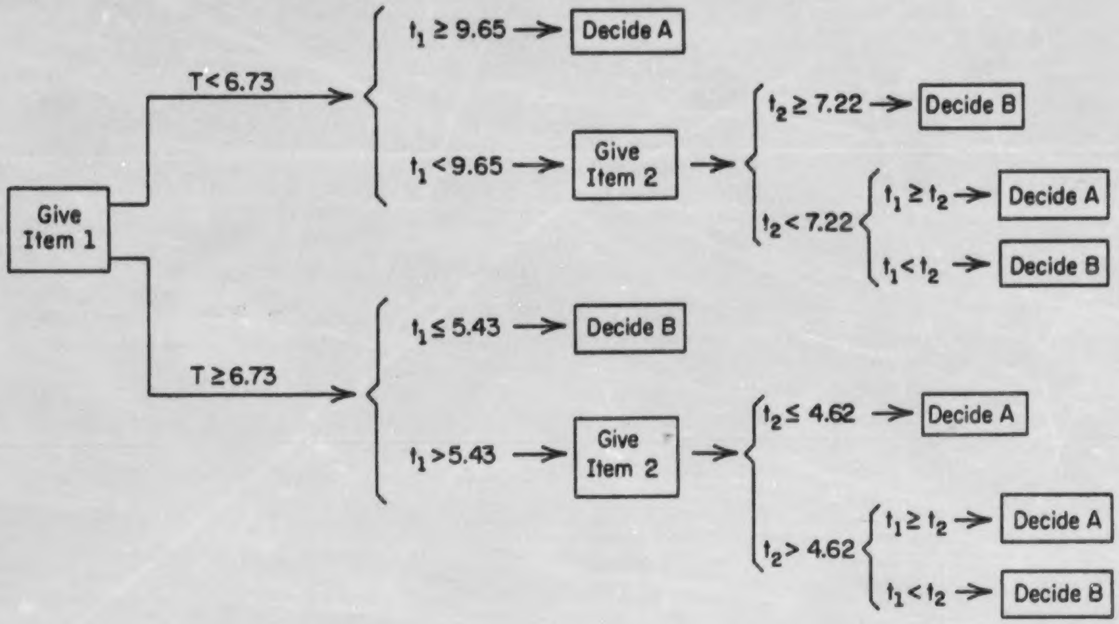
Let the means of Items 32 and 33 be taken as an example, then

t_{1A} = 9.65 sec., t_{1B} = 5.43 sec.

t_{2A} = 4.62 sec., t_{2B} = 7.22 sec.

Given these data and the observed response times t_1 and t_2 for the two items of a person about whom there is no other information, a natural but simple-minded decision rule for postdicting his/her instructional background would be to choose A if $t_1 > t_2$, and B otherwise. The problem, of course, is that the magnitude order of the two observed times could be reversed from the "true" order by errors of measurement. Knowledge of the person's τ_1 may help increase confidence in the postdiction, using the sequential decision rule shown in Figure 6, again a deliberately simple-minded one.

Figure 6
Sequential Decision for Postdicting Method A or B Based on Knowledge of T and Response Times for Items 1 and 2.


Give Item 1
T<6.73
t1 ≥ 9.65
Decide A
t1 < 9.65
Give Item 2
t2 ≥ 7.22
Decide B
t2 < 7.22
t1 ≥ t2
Decide A
t1 < t2
Decide B
T ≥ 6.73
t1 ≤ 5.43
Decide B
t1 > 5.43
Give Item 2
t2 ≤ 4.62
Decide A
t2 > 4.62
t1 ≥ t2
Decide A
t1 < t2
Decide B

First, only Item 1 is administered to this person. Now suppose his/her τ_1 is less than 6.73 sec. (the mean of the four mean response times listed above). Then, if $t_1 \geq 9.65$, A is chosen and testing is terminated. If, on the other hand, $t_1 < 9.65$, Item 2 is then administered, and B is chosen if $t_2 \geq 7.22$; otherwise, A or B is chosen accordingly as $t_1 > t_2$ or $t_1 < t_2$, respectively. When the person's τ_1 is greater than or equal to 6.73, the sequential decision will be the dual of the above. Namely, if $t_1 \leq 5.43$, B is chosen; if $t_1 > 5.43$, Item 2 is further administered, and A is chosen if $t_2 \leq 4.62$. If $t_2 > 4.62$, A or B is chosen according to the magnitude order of t_1 and t_2.

Refinements to this simple approach would include getting CRR distributions for each item, given instructional background and the value of τ. Further, with a suitable assumption concerning the distribution of τ, the posterior probability for each instructional background could be derived given τ and the magnitude for each order of t_1 and t_2. With more than two instructional backgrounds and a larger number of discriminating items, the magnitude order of two response times would be generalized to a vector of response times exhibiting different patterns, i.e., permutation of the magnitudes of the elements.

Person Conditional Response Rate (PCRR)

Some discussion is in order to explain why the Weibull family was chosen over the gamma, despite the latter's having a longer tradition of usage in response time models (e.g., Rasch, 1960; Restle & Davis, 1962). First, the gamma distributions are indicated when distinct stages are identifiable in the process of solving the tasks, in which case c must be an integer representing the number of stages. Second, the shape parameter c of the Weibull family has the interesting feature of apparently distinguishing between different information-processing skills associated with different instructional backgrounds. This feature is no doubt related to the fact that the magnitude of c (i.e., whether c is greater than, equal to, or less than 1) determines the nature of the item conditional response rate function (ICRR), which describes whether perseverance increases the chances of an examinee's responding to an item, whether responses occur at random times, or whether a point of diminishing returns is reached early. In other words, it can be said, as mentioned in the first section on the rationale of Weibull distributions, that c is sensitive to the degree of involvement students show. Two different instructional methods usually require different steps of information-processing skills; thus, each method requires a different degree of involvement in solving a given item. For example, some items in Table 1, such as "-11+10 = ?" in the signed number posttest, yield not only different values of c but also significantly different mean response times, depending on whether the sequential or number-lines method is used for answering, as dictated by the examinee's instructional background. Moreover, the convenient ICRR function is readily expressed in closed form for a Weibull distribution but cannot be so expressed for a gamma distribution, because the incomplete gamma function cannot be expressed analytically.

The ICRR function is the probability that an examinee who has not responded to an item by time t will do so within an infinitesimal time interval thereafter. When item response times follow a Weibull distribution, this function $H_g(t)$ is given by $f_g(t)/[1 - F_g(t)]$, where $f_g(t)$ and $F_g(t)$ are expressions Equations 1 and 2 with the parameters subscripted with a g for item g and, in this case, t_0 set equal to 0. Hence,

$$H_g(t) = c_g \, t^{c_g - 1} \Big/ u_g^{c_g - 1}. \qquad [6]$$

From this, the transition to PCRR is made in a manner analogous to going from an item characteristic curve (ICC) to a person characteristic curve (PCC), first suggested by Mosier (1940,1942), recently by Weiss (1973; Vale & Weiss, 1975) and discussed in greater detail by Lumsden (1978) and Trabin and Weiss

(1979). In their case, for each individual a plot is made of the proportions of items of varying difficulty (represented by the horizontal axis) that are passed by that individual. In the present case, however, the ordinate at each point along the horizontal axis representing the mean response time of an item would be the value of $H_g(\tau)$, where τ is the latent response time for the particular person, computed from Equation 6 with the parameter values proper to that item substituted for u_g and c_g. Note that when shape parameter c equals 1.0 for all items, the PCCR curves are identical for all persons. Thus, utilizing the negative exponential distribution (i.e., a special case of the Weibull distribution functions) for this purpose will not work.

Equation 6 defines a function whose curve characterizes the behavior of item g over time in terms of the probability of reaching an answer. The steeper the slope of a curve is, the greater the chance that item g will be solved as time elapses. The steepness of the curves is a characteristic attributed to a given item, similar to the item discrimination index in latent trait theory. A pseudo-CRR function on variable τ can be defined as follows:

$$H_{d_g}(\tau) = \frac{c_g}{u_g}\left(\frac{\tau + d_g}{u_g}\right)^{c_g - 1} \qquad [7]$$

Similarly, a pseudo-PCRR function is given as a function on a set of Equations 7, $H_{d_g}(\tau)$, $g=1,\ldots,n$. For a fixed person i,

$$G_i(H_{d_g}) = H_{d_g}(\tau_i) g=1, \ldots, n. \qquad [8]$$

It should be noted that the τ in Equations 7 and 8 is merely an arbitrary time point and bears no relation to a person's latent response time (except for coinciding in numerical value). Only in the context of the random variable $\tau + d_g$ does τ have the sense of latent response time; but to use $\tau + d_g$ as the argument of $H_{d_g}(\tau)$ would be meaningless, because $\tau + d_g$ is approximately the person's observed response time for item g, and it would be a contradiction in terms to speak of the person's responding in the next moment, given that he/she has not responded up to the actual time point at which he/she did respond.

The above remarks indicate that the particular approach attempted here for defining PCRR was futile, but not that the concept of PCRR itself is meaningless. An alternative, more justifiable approach might be to transform response time to an approximate normal variable. The transformation derived by the usual method of obtaining variance-stabilizing transformations was unusable because it was an arcsine transformation whose argument could exceed one.[1] Therefore, the usual method was extended by taking up to the second term, instead of only the first, in the Taylor series expansion on which the transformation is based. The result was that the transform y is the solution of the following rather formidable differential equation:

---

[1] This fact was noticed and pointed out by Jim Paulson at the 1979 Computerized Adaptive Testing Conference.

$$h(t)(y'')^2 + g(m)y''y' + f(m)y' = C, \qquad [9]$$

where

$$h(t) = \mu_0^4[\Gamma(1+4/c) - \Gamma^2(1+2/c)]/4 - [\mu_0^3\Gamma(1+3/c)(t-t_0)$$
$$- 2\mu_0^2\Gamma(1+2/c)(t-t_0)^2 + (t-t_0)^4]$$
$$g(t) = \mu_0^3\Gamma(1+3/c) - 3\mu_0^2\Gamma(1+2/c)(t-t_0) + 2(t-t_0)^3$$
$$f(t) = \mu_0^2\Gamma(1+2/c) - (t-t_0)^2$$

and c is an arbitrary positive constant.[2] If it is further assumed that τ is normally distributed (which seems reasonable by virtue of the central limit theorem, since τ_i is a person's mean response time over an infinite set of items, which may be regarded to exhibit local independence if unidimensionality holds), then y and τ would jointly follow a bivariate normal distribution. Hence, if their correlation ρ can be estimated (roughly analogous to communality estimation in factor analysis), the joint distribution would be uniquely determined. From this and the distribution of τ, the conditional distribution of y given τ can be determined. All quantities associated with persons having a particular τ value are computed from this conditional distribution.

Estimation of the Parameters

Interest is now in estimating c_g, u_g, and τ_i, $g=1,\ldots,n$, $i=1,\ldots,N$ simultaneously. The set of admissible values of these parameters must be chosen that makes the log-likelihood function, lnL, the maximum. Unlike the case of dealing with performance scores, response time represents two different cases--one is a group whose members obtained the correct answer and the other is a group whose members obtained the wrong answers. Response time in the OK subgroup means the time needed to attain a given goal using a successful information-processing skill (or skills); but it is not that simple with the NO subgroup. A brief investigation of error analysis for the NO subgroup indicates that various kinds of misconceptions at different progressive stages of reaching a correct answer for a given item might have occurred. Therefore, only the OK subgroup will be considered in this paper.

Differentiating the logarithm of Equation 5 by parameters c_g and u_g, respectively, and setting the results equal to zero gives the following simultaneous equations:

$$\frac{\partial lnL}{\partial c_g} = \sum_i \left[\frac{1}{c_g} + ln\left(\frac{\tau_i + d_{ig}}{u_g}\right)\left\{1 - \left(\frac{\tau_i + d_{ig}}{u_g}\right)^{c_g}\right\}\right] = 0 \quad , \qquad [10]$$

---

[2] This has not yet been solved but a mathematician colleague assures that it is soluble.

$$\frac{\partial lnL}{\partial u_g} = \sum_i \frac{c_g}{u_g}\left[\left(\frac{\tau_i+d_{ig}}{u_g}\right)^{c_g} - 1\right] = 0 , \qquad [11]$$

$$\frac{\partial lnL}{\partial \tau_i} = \sum_g \left[\frac{c_g-1}{\tau_i+d_{ig}} - \frac{c_g}{u_g}\left(\frac{\tau_i+d_{ig}}{u_g}\right)^{c_g-1}\right] = 0 . \qquad [12]$$

The maximum likelihood method using the Newton-Raphson iteration procedure provides estimates of the roots of Equations 10 and 11 where τ, i=1,...,N are substituted by the mean response time of each person over items, g=1,...,n. Then, the roots of Equation 12, after the newly estimated c_g and u_g values are substituted, are sought by the same procedure.

A sufficient, but not necessary, condition that any of these stationary values, u, c, be local maxima is that

$$\frac{\partial^2 lnL}{\partial c_g^2} = - \sum_i^{N_g}\left[\frac{1}{c_g^2} + \left\{ln\left(\frac{\tau_i+d_{ig}}{u_g}\right)\right\}\left(\frac{\tau_i+d_{ig}}{u_g}\right)^{c_g} ln\left(\frac{\tau_i+d_{ig}}{u_g}\right)\right] < 0 \qquad [13]$$

$$\frac{\partial^2 lnL}{\partial u_g^2} = - \frac{c_g}{u_g}\left[\sum_i^{N_g}\left\{\left(\frac{\tau_i+d_{ig}}{u_g}\right)^{c_g} -1\right\} + \sum_i^{N_g} c_g\left(\frac{\tau_i+d_{ig}}{u_g}\right)^{c_g}\right] < 0 \qquad [14]$$

$$\frac{\partial^2 lnL}{\partial \tau_i^2} = - \sum_g (c_g-1)\left[\frac{1}{(\tau_i+d_{ig})^2} + \frac{c_g}{u_g^2}\left(\frac{\tau_i+d_{ig}}{u_g}\right)^{c_g-2}\right] < 0 \qquad [15]$$

It should be noted that Equations 13 and 15 are always negative, but Equation 15 will be negative only when the estimates of u_g are close enough to be the roots of Equation 11. In some earlier stages of iterations, the condition for u_g to yield a local maximum might not be satisfied. Thus, it is important to select appropriate starting values for the estimates of u_g.

For estimation of τ, solutions of Equation 12 are sought for which Equation 15 is satisfied. If $c_g = 1$ for all g, $g=1,\ldots,n$, then Equation 12 becomes

$$\sum_g \left(\frac{1}{u_g}\right) = 0 \ .$$

Since scale parameter u_g equals the mean of observed response time when shape parameter c is equal to unity, this equation becomes equivalent to

$$\sum_g \left(\frac{1}{\bar{t}_g}\right) = 0 \ .$$

But the reciprocal of observed mean cannot be zero for any item g; therefore, there should be some g such that $c_g \neq 1$. This implies that the maximum likelihood method does not work for response time models associated with the negative exponential distributions as long as the models are formulated assuming unidimensionality. Moreover, the notion of PCRR, which is parallel to that of PCC, will not be applicable to these models. This is because CRR functions are always parallel to the horizontal axis when the occurrence of a response is a random event and all the random events are assumed to be of the same kind, which is the case of negative exponential distributions.

Numerical Example

The parameters in the model were successfully estimated[3] for sample data, the pretest data of the signed number arithmetic lessons. Unfortunately, the posttest data of signed number operations described above (see also Appendix Tables A-1, A-2, or A-3) could not provide stable estimates with this computer program. The sample size for Group 2 was too small and there was not a large enough number of items—the 14 items that were of most interest. When the observed response time data were fitted to the Weibull distributions before, it was observed that the items testing the same skill in the pretest showed a systematic change with the estimates of u and t_0 according to their order of presentation, even though the difficulties of these parallel items did not show any noticeable change.

With this new model the changes in the slopes of the parallel items have a strong tendency toward being monotonically increasing. For example, Items 3 and 45 ask "-3+2=?" and "-7+5=?," respectively. The dotted lines in Figure 7 are CRR functions associated with the observed response time, and the solid lines are the theoretically derived CRR functions. It is interesting to note that the random variable $\tau_i + d_{ig}$ can be rewritten as $(\tau_i - \bar{t}_{i.}) + t_{ig} = T_{ig}$. Denoting $\tau_i - \bar{t}_{i.} = -\varepsilon_i$, then T_{ig} can be expressed as $t_{ig} - \varepsilon_i$. Hence, the observed response time t_{ig} becomes the sum of a *true-score-like* T_{ig} and an "error" ε_i.

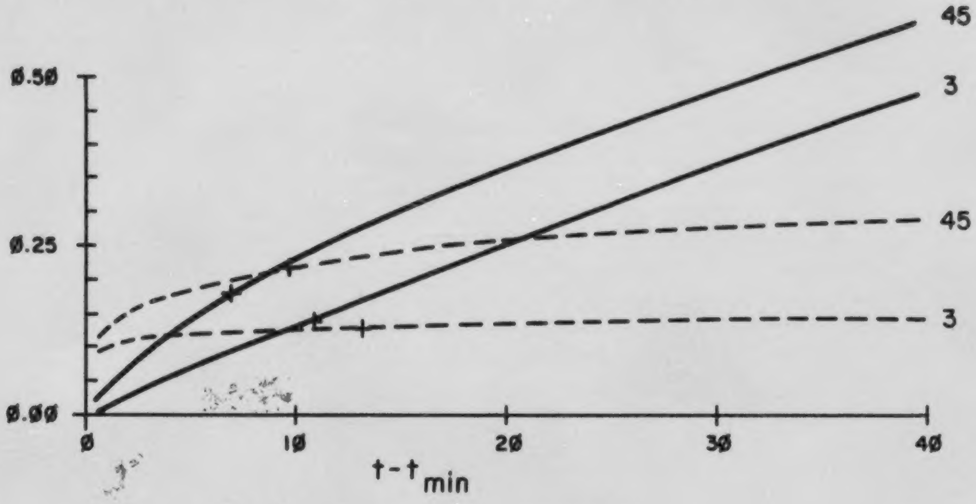
---

[3] A computer program for estimating parameters u_g, c_g, and τ_i for $g=1,\ldots,n$, $i=1,\ldots,N$ was written on the PLATO system by Robert Baillie.

Therefore it might be considered that the theoretical CRR functions are defined on a pseudo true score random variable, T_{ig}.

Figure 7
CRR Functions Defined on the Random Variable T_{ig}
$F'(t) \,/\, [\,1 - F(t)]$, $F(t)$ = Weibull Distribution Function

Item	t_0	c	μ_0	$\bar{t}$	
3	5	1.09	8.80	13.517	observed
45	4	1.22	6.28	9.883	
3	0	1.95	12.35	10.951	theoretical
45	0	1.68	8.36	7.465	



Summary and Discussion

The customary method for assigning a score to an individual on adaptive tests--or, for that matter, whenever a latent trait model is employed--is to use the estimate θ of the ability (or achievement) parameter. This may be adequate when the only purpose of testing is to calibrate the individual's ability or achievement level. When the further purpose of using θ as the basis for routing the student to a suitable starting point in a lesson series is involved, however, sole reliance on θ can create serious problems. This is because two examinees may have identical response patterns (and, hence, θ values) and yet differ drastically in the manners--the cognitive processes and information-processing skills that are brought into play--in which they arrived at their answers to the items, correct or incorrect, as the case may be. Efficient and effective routing of students to lessons requires this deeper diagnosis instead of mere information as to which items they get right or wrong.

Increasing recognition is being given to this fact, as evidenced by the number of studies either directly or indirectly germane to it that have recently been done by cognitive psychologists (e.g., Anderson, Kline, & Beasley, 1978; Carroll, 1978; Frederiksen, 1978; Greeno, 1977; Groen & Perkum, 1972; Heller &

Greeno, 1978; Rose, 1978; Sternberg, 1979). These studies have demonstrated the existence of a variety of cognitive processes, which differ from individual to individual.

One clue to the type of cognitive process employed by a student in solving a given problem can come from knowing his/her instructional background. Fortunately, a follow-up study of Tatsuoka and Birenbaum (1979) indicates that the Weibull shape parameter c, obtained by fitting response time data, is helpful to differentiate among various instructional methods associated with signed number operations. The Weibull distributions can be mathematically derived from the assumptions that the CRR--essentially, the conditional probability that a peson will respond to a given item during the interval $[t, t + dt]$, given that he/she has not responded to the item up to the time t--is monotonically increasing, decreasing, or constant. The slope of the CRR function for a given item is determined by the magnitude of the shape parameter and the mean of the item response time. If c is larger than 1, then CRR is a monotonically increasing function. If $c = 1$, then CRR is constant.

Some types of information-processing skill require a greater amount of involvement in a student's effort in solving a given problem, whereas others do not require so much to obtain the answer to the same item. The magnitude of the shape parameter c and mean response time for the former become noticeably larger than those for the latter. Therefore, the slopes of CRR functions differ in steepness to a greater extent. This sensitivity of the Weibull distributions to the procedures associated with different teaching methods is an advantage in dealing with psychological research. As Scheiblechner (1979) has stated, "the exponential or Weibull distribution is an adequate model for more sorts of psychological data than is commonly assumed if the parameteric structure of the latencies is properly chosen."

First, it was assumed that for a given set of items there exists a latent variable affecting the time taken by an examinee to answer each of these items. A model associated with response time, roughly paralleling latent trait theory, was formulated on the strength of the observed fact that the Weibull distribution fits the response time data for most items. The main concern in the model is to express the relationship between latent response time variable and the information-processing skills.

An estimation routine of the parameters by the maximum likelihood method was programmed and a numerical example was shown. The maximum likelihood method is not applicable to estimate Weibull parameters when all shape parameters are supposed to be 1, that is, the cases of negative exponential distributions. Further research will be necessary in exploring a different parameter estimation procedure, such as the conditional maximum likelihood method.

Information function of item g, $I_g(\theta)$ was integrated numerically and found to be always constant except for $c_g = 1$. However, its discussion will be reported in another paper.

The particular approach attempted here for defining item CRR and PCRR functions resulted in the loss of the attractive feature of capability to provide mathematical meaning to the curves in terms of τ_1. However, this attractive feature still holds for the variable mentioned in the numerical example, that is, T_{ig}. An alternative approach was outlined, but further research is necessary to make this approach operational.

REFERENCES

Anderson, J. R., Kline, P. J., & Beasley, C. M., Jr. A theory of the acquisition of cognitive skills. New Haven: Yale University, Department of Psychology, 1978.

Carroll, J. B. How shall we study individual differences in cognitive ability? —Methodological and theoretical perspectives (Technical Report 1). Chapel Hill: University of North Carolina, Psychometric Laboratory, 1978.

Frederiksen, J. R. Assessment of perceptual decoding and lexical skills and their relation to reading proficiency. In A. M. Lesgold, J. W. Pellegrino, S. Fokkema, & R. Glaser (Eds.), Cognitive psychology and instruction. New York: Plenum Press, 1978.

Greeno, J. G. Analysis of understanding in problem solving. Paper presented at the Developmental Models of Thinking Workshop, Kiel, West Germany, November 1977.

Groen, G. J., & Perkum, J. M. A chronometric analysis of simple addition. Psychological Review, 1972, 79, 329-343.

Heller, J. I., & Greeno, J. G. Information processing analyses of mathematical problem solving. Unpublished manuscript, University of Pittsburgh, 1978.

Lumsden, J. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 1978, 31, 19-26.

Mann, N. R., Schafer, R. E., & Singpurwalla, N. D. Methods for statistical analysis of reliability and life data. New York: John Wiley & Sons, 1974.

Mosier, C. I. Psychophysics and mental test theory: Fundamental postulates and elementary theorums. Psychological Review, 1940, 47, 355-366.

Mosier, C. I. Psychophysics and mental test theory II: The constant process. Psychological Review, 1941, 48, 235-249.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedogogiske Institut, 1960.

Restle, F., & Davis, J. H. Success and speed of problem-solving by individuals and groups. Psychological Review, 1962, 69, 520-536.

Rose, A. M. An information processing approach to performance assessment (Final Report). Washington, DC: American Institutes for Research, November 1978.

Sato, T. Diagnostic and formative evaluation of data processing system. In T. Sato (Ed.), Computer managed instruction system. Tokyo: Electronics Communication, 1977.

Scheiblechner, H. Specifically objective stochastic latency mechanisms. Journal of Mathematical Psychology, 1979, 19, 18-38.

Sternberg, R. J. New views on IQ's: A silent revolution of the 70s (Technical Report No. 17). New Haven: Yale University, Department of Psychology, April 1979.

Tatsuoka, K. K., & Birenbaum, M. The danger of relying solely on diagnostic adaptive testing when prior and subsequent instructional methods are different (CERL Report E-5). Urbana, IL: University of Illinois, Computer-based Educational Research Laboratory, March 1979.

Trabin, T. E., & Weiss, D. J. The person response curve: Fit of individuals to item characteristic curve models (Research Report 79-7). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1979.

Tatsuoka, K. K., & Tatsuoka, M. M. Time-score analysis in criterion-referenced tests (Final Report for NIE Project No. 6-0554, Grant No. NIE-G-76-0087). Washington, DC: U.S. Department of Health, Education and Welfare, National Institute of Education, February 1978.

Vale, C. D., & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, October 1975. (NTIS No. AD A018758)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis, MN: University of Minnesota Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376).

ACKNOWLEDGMENTS

This research was sponsored by the Personnel and Training Research Programs, Psychological Science Division, Office of Naval Research, under Contract No. N00014-78-C-0159, Contract Authority Identification Number, NR 150-415.

APPENDIX A: Supplementary Tables

Table A1
Kolmogorov-Smirnov Tests for the Total Sample on the Signed Number Test

Item No.	p	z	N	Item No.	p	z	N
1	.42	.88	59	33	.07	1.30	61
2	.97	.50	27	34	.38	.91	41
3	.15	1.13	55	35	.18	1.09	19
4	.51	.82	47	36	.39	.90	39
5	.01	1.71	68	37	.20	1.07	43
6	.98	.48	42	38	Not Tested		
7	.84	.62	25	39	Not Tested		
8	.29	.98	46	40	.21	1.06	78
9	.75	.68	37	41	.64	.74	47
10	Not Tested			42	.63	.75	76
11	Not Tested			43	.01	1.63	60
12	.13	1.17	87	44	.48	.84	31
13	.99	.43	24	45	.05	1.36	56
14	.69	.71	76	46	.37	.92	52
15	.18	1.10	59	47	.49	.83	69
16	.95	.52	25	48	.78	.66	34
17	.09	1.24	59	49	.86	.60	23
18	.15	1.14	60	50	.34	.94	25
19	.23	1.04	60	51	.65	.74	44
20	.87	.60	41	52	Not Tested		
21	.67	.72	20	53	Not Tested		
22	.92	.55	37	54	.03	1.44	86
23	.88	.59	33	55	.89	.58	28
24	Not Tested			56	.18	1.09	77
25	Not Tested			57	Not Tested		
26	.12	1.19	85	58	Not Tested		
27	.12	1.19	44	59	.00	1.75	80
28	.79	.65	71	60	.66	.73	21
29	.01	1.71	57	61	Not Tested		
30	.59	.77	25	62	Not Tested		
31	.23	1.04	58	63	.00	2.22	83
32	.19	1.08	62	64	.42	.88	46

Table A2
Kolmogorov-Smirnov Tests for Group 2 on the Signed Number Test

Item No.	p	z	N	Item No.	p	z	N
1	.98	.47	16	33	.85	.61	20
2	1.00	.26	4	34	1.00	.34	7
3	.95	.53	18	35	.00	.00	0
4	.76	.67	13	36	.97	.49	10
5	.42	.88	23	37	.87	.59	9
6	.99	.45	10	38	Not Tested		
7	1.00	.29	3	39	Not Tested		
8	1.00	.39	10	40	.75	.67	26
9	.76	.67	8	41	.77	.66	14
10	Not Tested			42	.98	.47	24
11	Not Tested			43	.52	.81	16
12	.61	.76	26	44	1.00	.35	2
13	.86	.61	7	45	.78	.66	18
14	.91	.56	24	46	.85	.61	18
15	.96	.51	16	47	.25	1.01	23
16	1.00	.29	3	48	.52	.81	9
17	.98	.46	18	49	1.00	.35	2
18	.83	.63	20	50	.98	.47	4
19	.71	.70	19	51	.70	.71	7
20	1.00	.35	8	52	Not Tested		
21	1.00	.35	2	53	Not Tested		
22	1.00	.41	6	54	.38	.91	26
23	.99	.44	7	55	.89	.58	3
24	Not Tested			56	.67	.73	25
25	Not Tested			57	Not Tested		
26	.72	.69	26	58	Not Tested		
27	.94	.54	11	59	.85	.61	25
28	.88	.59	23	60	1.00	.35	2
29	.91	.56	16	61	Not Tested		
30	1.00	.29	3	62	Not Tested		
31	.90	.57	15	63	.88	.59	26
32	1.00	.38	19	64	.94	.54	15

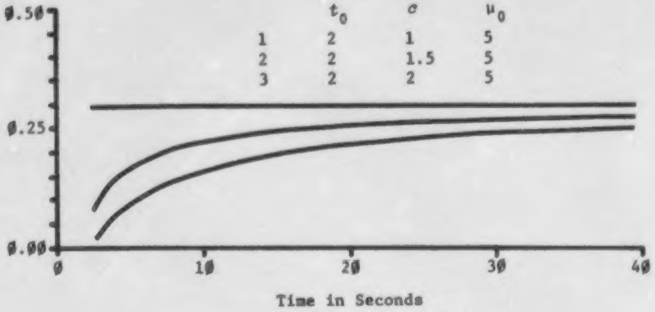
Table A3
Kolmogorov-Smirnov Tests for the Group other than Group 2 on the Signed Number Test

Item No.	p	z	N	Item No.	p	z	N
1	.45	.86	43	33	.59	.77	41
2	.94	.53	23	34	.37	.92	34
3	.32	.95	38	35	.18	1.09	19
4	.82	.63	34	36	.36	.92	29
5	.10	1.22	45	37	.69	.71	34
6	.99	.44	32	38	Not Tested		
7	.93	.54	22	39	Not Tested		
8	.89	.58	35	40	.16	1.13	52
9	.99	.46	29	41	.93	.54	32
10	Not Tested			42	.92	.55	52
11	Not Tested			43	.06	1.32	44
12	.17	1.11	61	44	.63	.75	28
13	.99	.45	17	45	.14	1.15	38
14	.49	.83	52	46	.59	.77	34
15	.64	.74	43	47	.74	.68	47
16	.97	.49	22	48	.62	.76	25
17	.59	.77	40	49	.90	.57	21
18	.31	.97	40	50	.35	.93	21
19	.49	.83	40	51	.63	.75	37
20	.86	.60	33	52	Not Tested		
21	.43	.87	18	53	Not Tested		
22	.88	.59	32	54	.07	1.28	60
23	.80	.64	26	55	.99	.43	25
24	Not Tested			56	.24	1.03	52
25	Not Tested			57	Not Tested		
26	.19	1.09	59	58	Not Tested		
27	.23	1.04	33	59	.05	1.37	55
28	.95	.52	48	60	.72	.69	19
29	.03	1.45	41	61	Not Tested		
30	.85	.61	22	62	Not Tested		
31	.40	.89	43	63	.00	2.04	57
32	.26	1.01	43	64	.69	.71	31

APPENDIX B: Supplementary Figure

Conditional Response Rate of Three-Parameter Gamma Distribution

$F'(t)/[1-F(t)]$, $F(t)$ = Gamma Distribution Function



LATENT TRAIT SCORING OF TIMED ABILITY TESTS

DAVID THISSEN
UNIVERSITY OF KANSAS

The advent of computerized testing has made timed testing a feasible process. Paper-and-pencil testing technology limited the test theorist to an analysis of the responses of the examinees. Computerized testing, on the other hand, has the potential to provide the tester with a great deal more information than that contained in the responses alone. As adaptive tests become more efficient, and as they become shorter, each item and its associated response must provide more and more information about the ability of the examinee. But only a limited amount of information can be obtained from binary responses, and even the use of three or more response categories provides only limited increases in the amount of information provided by any one item (Bock, 1972; Samejima, 1969; Thissen, 1967b). The additional information needed must come from some other response variable, preferably a continuous one; response latency is a likely candidate.

Although there exist completely general latent trait test item response models for ordered or for strictly nominal item responses in any number of response categories (Bock, 1972; Samejima, 1969), as well as at least one relatively specialized latent trait model for continuous item responses (Samejima, 1973), there has been little work on item response models for timed test data. White (1973) proposed a model for individual differences in speed and accuracy in a timed testing situation; but in that system the response latencies were used as a fixed part of the model rather than as observations measured with error. It seems more in keeping with the spirit of latent trait test theory to consider the latencies, like the item responses themselves, to be fallible data reflecting underlying trait values.

In research predating contemporary statistical estimation schemes for the application of latent trait item response models, Furneaux (1961) suggested that the normal ogive model would serve quite well for item responses obtained in a timed testing situation. He found that for timed responses to letter series problems, the log transformed latencies were linearly related to heuristically estimated normal ogive ability and difficulty parameters. Following lines suggested by Furneaux's findings, Thissen (1976a) suggested an integrated model for the item responses and latencies obtained in timed testing and developed a scheme for the joint maximum likelihood estimation of the parameters of that model.

In that earlier research, the results of the application of the timed testing model to data from a test of spatial visualizing ability, as well as to data from the Matching Familiar Figures test (Kagen, Rosman, Day, Albert, & Phillips, 1964) and to a laboratory perception task, were discussed. The present paper concerns the application of the same timed testing model to data obtained with three tests of classical form: a subset of the Raven (1956) Progressive Matrices, a version of the Guilford-Zimmerman (1953) Aptitude Survey "Clocks" Spatial Visualization Test, and a Verbal Analogies test drawn from the Minnesota Multimodal Analogy Test, consisting of items described by Tinsley (1971) and analyzed extensively by Whitely (1977).

The Model

The item response model used here was actually developed primarily on the basis of exploratory data analysis of timed test data; that development is described elsewhere (Thissen, 1976a). In this section a development of this model, which is not entirely based on the form of timed test response data, will be discussed.

The technology for latent trait item analysis of item response data is well established when the items are scored dichotomously. The logistic model described by Birnbaum (1968) has been useful in many situations; therefore, it would seem appropriate to use that model, at least as a first approximation, for the item response data obtained in the timed testing situation. If the item responses r_{ij} for person i responding to item j are r_{ij} = 1 and if the response of person i to item j is correct and 0 otherwise, then the logistic model is

$$P(r_{ij} = 1) = 1/[1 + \exp(-z_{ij})], \qquad [1]$$

where

$$z_{ij} = a_j\theta_i + c_j;$$

and

$$P(r_{ij} = 0) = 1 - P(r_{ij} = 1). \qquad [2]$$

In the context of the timed testing situation, θ_i could be called the "effective ability" of person i (following White, 1973) to distinguish this ability estimate--which is obtained under circumstances allowing the examinee any amount of time to respond--from more conventional ability estimates obtained with speeded tests. The parameter a_j is the discrimination parameter, or "slope," of item j; it reflects the (possibly) different degrees of relationship between the items and the trait being measured. The parameter c_j is the easiness of item j.

To follow the traditional forms of latent trait test theory, the response time of person i on item j, t_{ij}, should also be a function of some parameters describing characteristics of person i and item j. It would be interesting and useful if those parameters were the same as the parameters that are used to describe the item responses; that would mean that the variations in item responses and response times were attributable to the same sources and that the responses and latencies could both be used in the measurement of the ability of each examinee and the easiness of each item. Latency should clearly be a decreasing function of the $\underline{z}_{ij}$; that is, increases in $\underline{z}_{ij}$ (which imply increased person ability or item easiness) should be related to decreases in response latency. The form of the random error involved in the measurement of response latency must also be specified; data analytic considerations suggest that response times are frequently approximately log-normally distributed, so a linear model for the logarithm of latency could be assumed to include Gaussian error. This suggests the following model:

$$\log(t_{ij}) = v - bz_{ij} + \varepsilon_{ij};\ \varepsilon_{ij} \sim N(0,\sigma^2)\ , \qquad [3]$$

in which $\underline{v}$ is the overall mean log response time and $\underline{b}$ is a regression parameter reflecting the relationship of effective ability and easiness (in the logit $\underline{z}_{ij}$) with latency, and the scale conversion of logits to log seconds. Of course, the parameters of $\underline{z}_{ij}$ in Equation 3 are the same as the parameters of $\underline{z}_{ij}$ in Equation 1; they may be simultaneously estimated using current constrained maximum likelihood estimation techniques.

However, as it stands, the model described by Equation 3 is unlikely to provide a very good fit to observed data. There are likely to be attributes of both the examinees and the items that contribute consistently to latency but that are unrelated to the trait the items are intended to measure. For verbal items either the length of the item (and the associated reading time) or the number of attempts required to obtain the needed semantic data may contribute to response latency; but these factors could be unrelated to the easiness of that item (and the probability of a correct response). Some examinees may press the response keys more slowly than others in a pattern that is consistent but unrelated to their ability. The addition of person and item slowness parameters, $\underline{s}_i$ and $\underline{u}_j$, to Equation 3 results in the model

$$\log(t_{ij}) = v + s_i + u_j - bz_{ij} + \varepsilon_{ij}, \qquad [4]$$

where

$$\varepsilon_{ij} \sim N(0,\sigma^2).$$

The hybrid model formed by the combination of the model in Equation 1 for the item response data and Equation 4 for the response times will be used in the analysis of timed test data in the sequel. This hybrid model includes a number of assumptions about functional forms: The logistic is used as the probability of a correct item response as a function of effective ability, and additive linear effects and Gaussian error are assumed for the log of response time. Assumptions such as these must be made to provide the basis for the maximum likelihood parameter estimation.

Two comments are in order about the functions included in the model. First, after the parameters of the model have been estimated, the appropriateness of the assumed functional forms may be evaluated to some extent by an examination of the goodness of fit of the various features of the model. The empirical proportions of correct responses at various levels of ability may be compared to the logistic predictions; the normality of the actual residuals of log(t) from the linear model may be evaluated. To the extent that the functional forms included in the model follow the observed data, the parameters of the model should provide a useful summary of the item response data.

Secondly, this model is not being proposed as a process model for the psychological description of item response behavior. Although such process models have been developed for choice behavior (see Audley, 1960, 1973; Laming, 1968; Luce, 1960) and are currently being developed for specific types of complex cognitive abilities (see Sternberg, 1977; Whitely, 1979), they tend to be too complex for current application in practical testing and measurement. The model proposed here follows the tradition of earlier test theory in that it attempts to generally describe responses to a variety of possible kinds of test items. It is to be hoped that the parameters of the hybrid timed testing model may be useful in some psychological process research, such as summaries of attributes of test items and examinees, which may be compared to theoretical predictions. Beyond that, the current model is meant to be a general model for the measurement of examinees and the calibration of test items; such a model will likely be superseded by accurate, complete psychological models for cognitive processes as soon as they become available for cognitive tasks in the domain of ability measurement.

Estimation of Model Parameters

The parameters of the timed testing model to be estimated form the set $\xi = \{\theta, s, a, c, u, v, b, \sigma^2\}$; there are $2N + 3n$ independent parameters (for N persons and n items), since the location and scale of θ, as well as the location of the slowness parameters, must be set arbitrarily. With the usual assumption of local independence extended to include the assumption that the error of observation of latency is independent of the response to the same item, conditional on the parameters, the likelihood of the entire set of data, given the parameters may be written

$$L(R,T|\xi) = \Pi_i \, \Pi_j \, P_{ij} \, H_{ij} \tag*{[5]}$$

in which

$$R = [r_{ij}],$$

$$T = [t_{ij}],$$

$$P_{ij} = P(r_{ij} = 1)^{r_{ij}} \, P(r_{ij} = 0)^{(1 - r_{ij})} ,$$

and

$$H_{ij} = \Phi(d_{ij}) ,$$

where

$\phi(d)$ is the standard normal density

and

$$d_{ij} = \{\log(t_{ij}) - [\nu + s_i + u_j - bz_{ij}]\}/\sigma \ .$$

A two-stage procedure for locating the maximum of $L(R, T|\xi)$ is described in detail by Thissen (1976a). The first stage of the procedure makes use of the conjugate gradient method (Fletcher & Reeves, 1964) for function minimization (maximization, in this case) to approach the joint maximum of the log likelihood in the 2N + 3n dimensional parameter space. This procedure requires 1.5 to 2 times as many iterations as there are parameters to be estimated; but it proceeds quickly, as the conjugate gradient technique requires computation only of the first derivatives of the log likelihood. As the maximum is approached, and the corrections become smaller, the conjugate gradient algorithm generally encounters difficulty in locating an appropriate direction in the parameter space, which has dimensionality of several hundred. At that point the second stage of the estimation procedure is entered. The second stage is a cyclic procedure in which each person's parameters are estimated individually using the current values of the item and the overall test parameters; then each item's parameters are similarly estimated, the overall test parameters are revised, and the procedure is repeated. In the second stage, the maximizations in few dimensions for each subject and item are performed using a conditioned Newton-Raphson algorithm programmed by Haberman (1970). By the time the Newton-Raphson corrections are of the order 10^{-3} for each parameter, there is generally no appreciable change between cycles, and the procedure is stopped.

A useful feature of the second stage estimation procedure is that since the Newton-Raphson iterations require matrices of second partial derivatives for the person parameters (within each person) and the item parameters (within each person) and the item parameters (within each item), those matrices may be treated as information matrices and may be inverted to give estimated standard errors for the parameters. The resulting estimates for the standard errors are, unfortunately, somewhat smaller than they ought to be, as their computation ignores the fact that all of the other person (or item) parameters have also been estimated from the same set of fallible data. However, very limited monte carlo results (Thissen, 1976a) indicate that the bias is not too large.

The Data

The data analyzed were the responses of 78 University of Kansas undergraduate students to three tests of cognitive ability. The students were selected to participate in a study of individual differences in cerebral laterality; therefore, left-handed individuals were substantially over-represented in the sample. The students partially satisfied a research participation requirement in an introductory psychology course by their participation in the study. The students were largely college freshman: 36 were male and 42 were female.

Tests

The Verbal Analogies test consisted of a 27-item subset of the 60 verbal analogies from the Minnesota Multimodal Analogies test described by Tinsley (1971) and Whitely (1977). These analogies are quite difficult for college students; the raw scores ranged from 4 to 24, and the average proportion correct was .57. No item was answered correctly by all of the students.

The Progressive Matrices Test consisted of Sets B, C and D of the Raven Colored (Set B) and Standard (Sets C and D) Progressive Matrices (1956). There were thus a total of 36 items, of which two (B-1 and B-3) were answered correctly by more than 98% of the students; those two items were omitted from further analyses. The Progressive Matrices Test was easier for this group than the Verbal Analogies test; the average proportion correct was .72. The raw scores ranged from 12 to 34.

The Clocks spatial test consisted of 19 items drawn from a set of color slides made to resemble items on the Guilford-Zimmerman (1953) "Clocks" test of spatial visualizing ability. The items give pictorial instructions about the rotation in three dimensions of a large, old-fashioned alarm clock, and the examinee is required to select from a set of alternatives a view of the clock as it would appear in the rotations. Two of the items were too easy for this group and were omitted from the analyses; therefore, Clocks was a 17-item test, on which the raw scores ranged from 5 to 16.

The test items were presented by a slide projector and rear projection screen, which was located immediately in front of the examinee. All test items were in multiple-choice format, with from four to eight alternatives; the students responded by pressing a numbered response key on a calculator-like keyboard. The presentation of the slide triggered a light sensor, which started the timer; the response of the examinee stopped the timer and initiated a display of the response and latency for the examiner.

Data Preparation

Before the iterative estimation of the parameters of the latent trait model was begun, the data were trimmed of extreme outliers, following the procedure suggested earlier (Thissen, 1976a). Using heuristically obtained starting values for the parameters of the timed testing model, observations for which the observed latency deviated from the predicted latency of the model by more than three standard deviations were removed from the sample. Since missing data are tolerated by the estimation procedure, those latencies were not replaced in any way. Most responses trimmed had very long latencies (more than a minute in many cases). Less than .5% of the observations were trimmed in this way.

Results

Verbal Analogies

The Verbal Analogies included in this set were unconventional in a variety of ways: The blank (to be filled in by the examinee's response) could be any one of the four terms of the analogy; the terms in the analogy could be permuted (from a:b::c:d to a:c::b:d, and so on); and a number of different types of relationships could be obtained between the analogy elements. All of these factors apparently contributed to complaints that it was a very bad test from both the research assistants who administered the test and the students who took the test. That it was a very difficult test also probably contributed to the negative feelings the examinees voiced. Nevertheless, the latent trait analysis with the timed testing model revealed that the response data from this test were fitted by the model more closely than any other test data to which the model has been applied. And the results which were obtained in the comparisons of the results of this analysis to previous analyses of this same pool of items indicates that this was, in fact, a very good test for measuring analogical reasoning ability, even though the examinees did not like it.

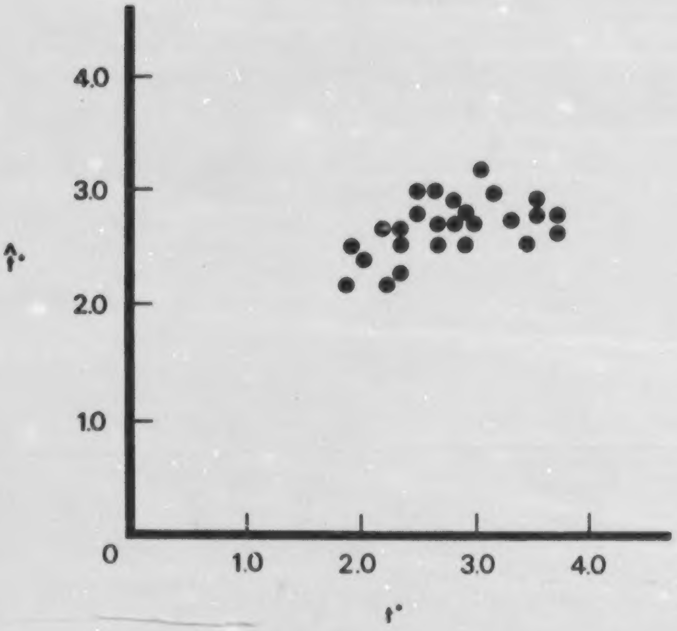
The goodness of fit of the data to the timed testing model may be measured in a number of ways. Since the logistic item response model gives a probability of a correct response at any value of θ, the examinees may be ordered according to their estimated effective ability level, divided into relatively homogeneous ability groups, and the proportion of correct responses in each of the ability groups may be compared to the proportion expected if all of the examinees in each group had the same ability, namely, the mean for that ability group. This procedure yields a contingency table (correct/incorrect responses × ability fractile) for each item and a χ^2 test of the goodness of fit of the model. In this analysis the examinees were divided into 6 equal groups and the resulting 4 degree-of-freedom χ^2 tests were computed; only 1 of the 27 values was significant at the $p = .05$ level. The total of the likelihood ratio χ^2 values was 131.2 with 108 degrees of freedom ($.10 > p > .05$), which was nonsignificant and indicates that the probabilities of correct responses are predicted fairly well by the model.

Once the parameters of the model were estimated, the assumption of normality of the residuals of log (t) from the model were examined by computing the actual residuals (2,098 of them, in this case), dividing them into fractiles, and comparing their distribution to Gaussian expectation. Again, a χ^2 test of the goodness of fit was used; in this example, for 10 fractiles normal error seemed to be met in the data. The result masks a small violation of the assumptions of the model: Correct responses were slightly, but significantly, faster than incorrect responses; the difference was .22 standard deviation units in log time.

The goodness of fit of the model for the latencies may be strikingly portrayed using data from individual examinees and items. Figure 1 shows a scatterplot of the observed log response times, t^*, and the log response times estimated from the model, $\hat{t}^*$, for an individual student across the 27 items. The correlation was .59.

Figure 2 is a similar scatterplot of the log latencies, t^*, and the predicted latencies, $\hat{t}^*$, for all 78 subjects on one of the analogy items, "cent:dime::......:dollar." The correlation between the observed and fitted log latencies for this item was .65.

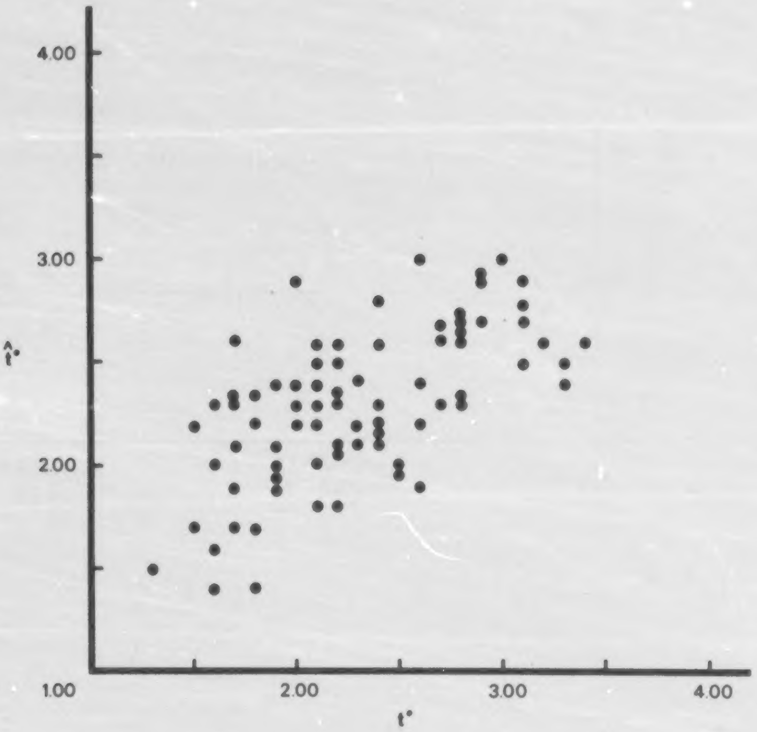
Figure 1
Scatterplot of the Observed Log Response Times, $\underline{t}^*$, and the Fitted Log Response Times, $\hat{\underline{t}}^*$, for Student 24 for the Analogies Test



The fit to the latency data comes both from the person and item slowness parameters, $\underline{s}_i$ and $\underline{u}_j$, and from negative effects due to effective ability and item easiness, since the logit regression parameter, $\underline{b}$, had a value of .20. The quantitative meaning of the regression of latency on effective ability varies along the time scale due to the log transformation; but for an average item (in all respects) a person of average slowness and average effective ability ($\theta = 0$) should respond in about 12.5 seconds and a less able individual ($\theta = -1$) should respond in about 15.2 seconds. The timed testing model makes use of the relationships between latency and the probability of a correct response to provide relatively precise estimates of all of the parameters involved. Some 35% to 50% of the variance in log latency within an individual or an item is predicted by the model.

Since some of the item parameters in the timed testing model are the same as those used in conventional latent trait analysis, the timed testing model may be evaluated as an item calibration scheme by comparing the item parameters estimated using the timed testing model with item parameters estimated using large-sample, response-only latent trait techniques. Item easiness was estimated by Tinsley (1971) for the analogy items used here with the Rasch (1960) 1-parameter logistic model, with data from 641 subjects. In Figure 3 the resulting item easiness parameters are plotted against $\underline{c}_j/\underline{a}_j$, the corresponding transformation of the easiness parameter of the 2-parameter logistic model used here. The correlation between the two sets of easiness parameters was .80 (or .70 when the

Figure 2
Scatterplot of the Observed Log Response Times, t*, and the Fitted Log Response Times, $\hat{t}$*, for One Analogy Item



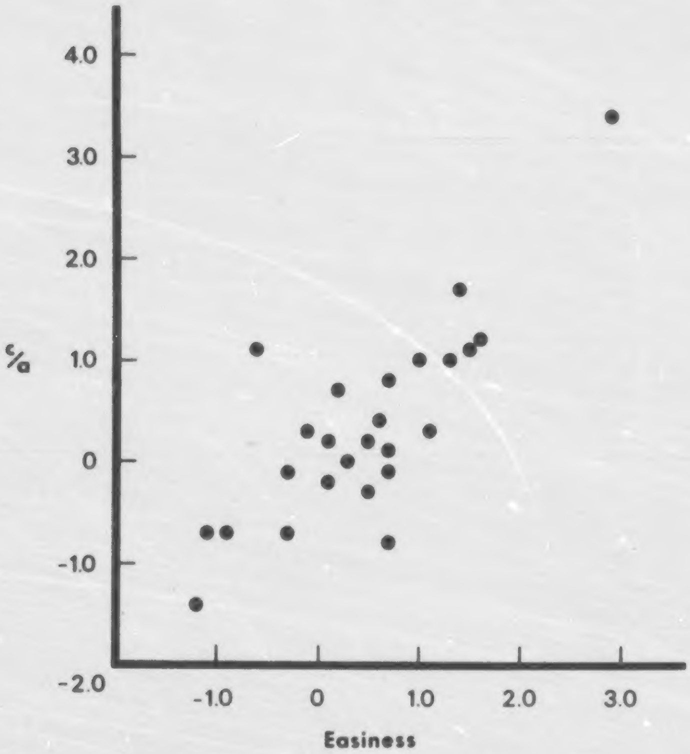
outlier in the upper right-hand corner, which was the easiest item in the test, was removed).

Large-sample estimates of 2-parameter logistic slopes were not available for this group of items. However, slopes for each item estimated using the 1-parameter logistic ability were available from the same large-sample item calibration study, and in Figure 4 they are plotted against the $\underline{a}_j$ estimated in the present study. The correlation was only .50, and only two-thirds of the points were within two of their own standard errors from the regression line.

The results shown in Figure 4 do not seem very good, until it is recalled that the large-sample slope estimates were not very precise, since they were estimated using ability derived from a 1-parameter logistic model; and, even with the timed testing model, the standard errors of slopes estimated with only 78 subjects were liberally estimated to be about 0.12 for this test. All things considered, the relationship between the large-sample slopes and those estimated by the timed testing model was about as strong as would be expected, given the estimation procedures and sample sizes involved.

The item parameters estimated in this analysis revealed some construct and

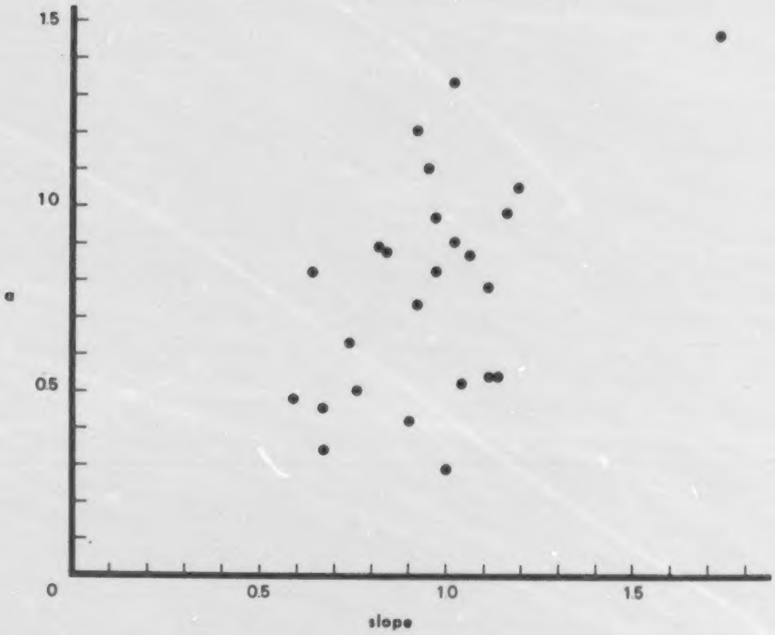
Figure 3
Timed Testing Easiness ($\underline{c}_j/\underline{a}_j$)
Plotted Against Large-Sample Easiness Parameters
for the Analogies Items



face validity when they were related to individual item characteristics. Using data from 107 persons (who sorted these and other analogy items into categories based on perceived similarity) and Wiley's (1967) latent partition analysis, Whitely placed these analogy items in 8 categories defined by the type of relationship between the elements. The average item parameters for the 8 categories of analogies are given in Table 1. The easiness and item slowness parameters did not vary significantly among the categories, but the discrimination parameters were significantly related to analogy type (F (7,18) = 3.16, $\underline{p} < 0.05$). Quantitative analogies (e.g., cent:dime::......:dollar), word pattern analogies (e.g., owl:ant::....:tan), and functional analogies (e.g., tree: man::sap:....) were strongly related to the analogy-solving ability being estimated. Class-naming and similarity analogies (e.g., puzzle:......::riddle:ocean), which are thinly disguised vocabulary items, were not strongly discriminating as analogies.

Within analogy types and within items of the same easiness, the $\underline{u}_j$ parameters described other differences among the items. Some 58% of the examinees responded correctly to the quantitative analogy "cent:dime::......:dollar" in a median time of 9.4 seconds, with an estimated item slowness of -.13. Nearly the

Figure 4
Timed Testing Discrimination Parameters ($\underline{a}_j$)
Plotted Against Large-Sample Slope Parameters
for the Analogies Items



same number, 59% of the students, responded correctly to another of the quantitative analogies, "......:yesterday::tomorrow:today"; but since they expended an average of 15.5 seconds on it, $\underline{u}_j$ was .28.

Is it possible that students could compute money faster than time? Or is a rearrangement of the second analogy in order to place the stem first the time-consuming factor? In any event, the consequences of this are that an examinee who responds to the first analogy correctly in 10 seconds is average; but an

Table 1
Mean Item Parameters for Eight
Categories of Analogies

Analogy Category	$\underline{a}$	$\underline{c}_j/\underline{a}_j$	$\underline{u}_j$
Quantitative	1.16	.46	.08
Word Pattern	.91	.21	.01
Functional	.90	.02	-.23
Opposites	.80	.73	.14
Conversion	.77	.59	.03
Class Membership	.67	1.31	.12
Class Naming	.53	-.47	-.26
Similarities	.42	.05	.09

examinee who responds to the second analogy correctly in 10 seconds is either very fast or very able. This information will be required by a computerized adaptive testing system, which attempts to use latency as part of a system to estimate the ability of examinees. And if it can be determined why "yesterdays" take longer than "cents," the result may contribute to the psychological understanding of analogy items.

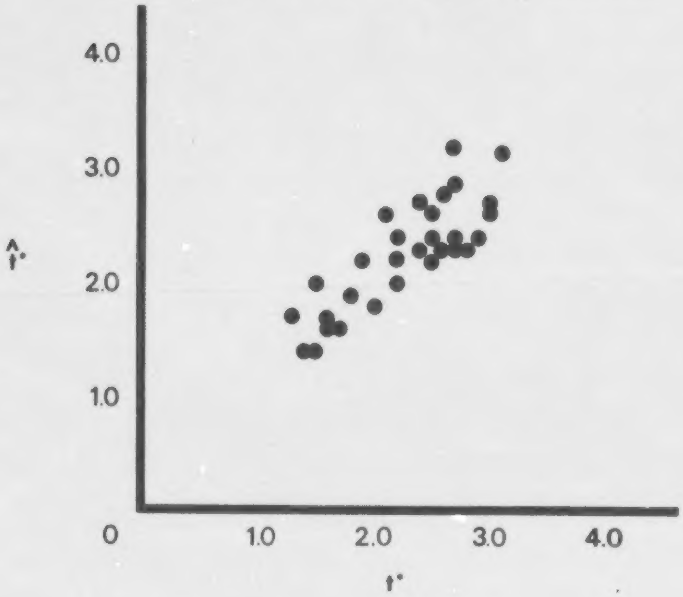
Progressive Matrices

The goodness of fit of the Progressive Matrices data to the timed testing model was not as good as for the Verbal Analogies. When the observed proportions of correct responses for a similar set of six ability groups were compared to the estimated logistic proportions of correct responses, 4 of the 34 individual item χ^2 tests were significant at the $p < .05$ level, and the total of the χ^2's was 201.0 with 136 df ($p < .01$), which indicates an overall significant lack of fit of the model. However, the problem seemed to be the test items, not the timed testing model, as the significant χ^2's were all due to items for which the observed proportions of correct responses did not form a strictly monotonic increasing function over estimated ability. For most of the items, the fit was acceptable; so the significant, but small, lack of fit of the item model did not seem to present insurmountable problems.

The χ^2 test of the goodness of fit of the Gaussian distribution to the log latency residuals was highly significant: The likelihood ratio χ^2 was 58.5, with 7 df ($p < .01$). This indicated a substantial violation of the assumption of lognormal error for the response times. The problem appeared to arise from two sources; the major problem arose because in this test, as in the Verbal Analogies, correct responses were faster than incorrect responses (even after all of the model corrections for ability, easiness, and so on). But the Progressive Matrices were much easier than the Verbal Analogies, so almost three-fourths of the responses were correct, and a little too fast. The combined effects of the 75 to 25 mixture of correct and incorrect responses and their slightly different error distributions made the total distribution of errors around the log latency model somewhat skewed, and that is what the goodness-of-fit test was detecting. This slight skewness (magnified in the χ^2 statistics by the 2,646 residuals in the distribution) was mostly in the middle of the distribution and should have little effect on the parameter estimates. There was also a 1% surplus of latencies that had long positive residuals (over 2 SD above the mean); but none of those were such outliers that they would exert excessive influence on the parameters.

Oddly enough, given those results from the goodness-of-fit statistics, the prediction of the log latencies from the timed testing model was better for the Matrices than it was for the Verbal Analogies. Figure 5 is a scatterplot of observed ($\underline{t}^*$) and fitted ($\hat{\underline{t}}^*$) log latencies for the 34 Matrices items for the same student whose Analogies data is in Figure 1. The correlation was .84, indicating that some 70% of the variance of log response time within that individual was being captured by the model. In part, this is because there was substantial regression of log latency on effective ability and easiness for the Matrices data. The b parameter was estimated at .8, and the location parameter for the log latencies was 3.10; that means that an item of easiness "0" should take an average examinee 22.2 seconds, and an item of easiness "1" would take the same examinee only 10 seconds. It also means that the response times were used very heavily in the estimation of effective ability and item easiness.

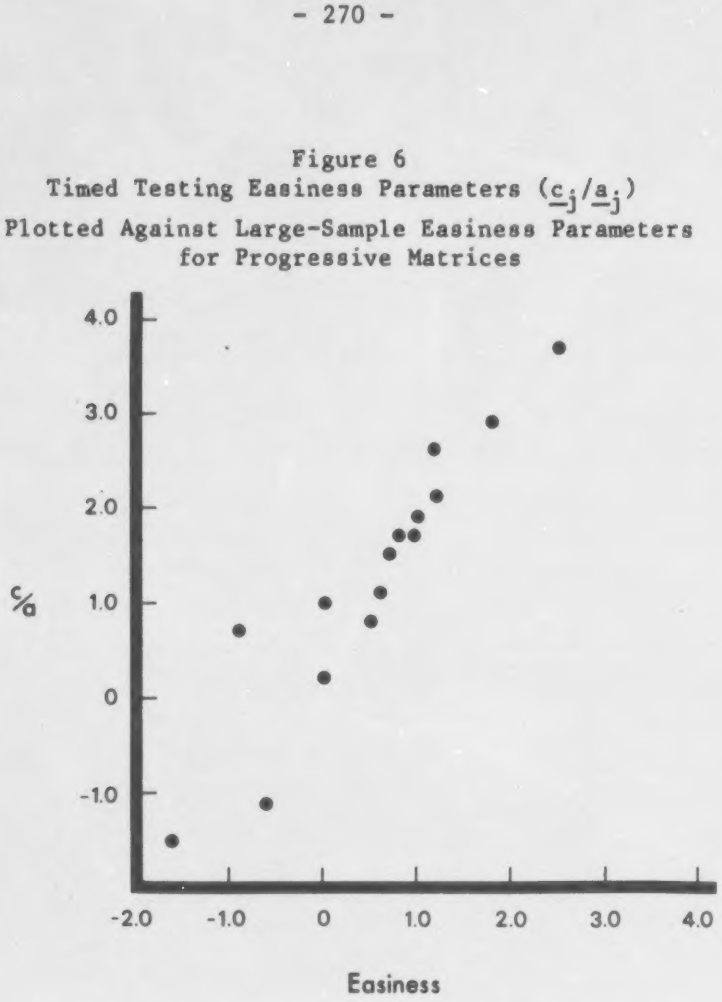
Figure 5
Scatterplot of the Observed Log Response Times, $\underline{t}^*$,
and the Fitted Log Response Times, $\hat{\underline{t}}^*$,
for Student 24 for Progressive Matrices



Some evidence that the item easiness and slope parameters were estimated fairly well by the timed testing model with the aid of the response times comes from a comparison of the item parameters obtained in this analysis with those of another large-scale latent trait item calibration that included a subset of these same Progressive Matrices items. Thissen (1976b) estimated 2-parameter logistic item parameters for 20 items drawn from the Progressive Matrices Sets A, B, and C administered to 570 junior high school students. Eighteen of those items formed a subset of the 34 Matrices items included in the present analysis.

Figure 6 is a plot of the corrected easiness parameters ($\underline{c}_j/\underline{a}_j$) from the present data against the same parameter for the same items from the junior high data; the correlation was .94. Figure 7 is a similar scatterplot for the slopes ($\underline{a}_j$). The correlation in this case was .61, and there was some evidence of curvilinearity, due mostly to the fact that there were 4 very high slopes in the earlier junior high calibration with only 20 items. (Frequently, one or a few of the slopes "climb" in a 2-parameter logistic item calibration of a test with few items. This did not occur with this relatively large set of 34 Matrices items.) Nevertheless, these sets of discrimination parameters correlated more highly than the unmatched slope parameters did for the Verbal Analogies. The correlation indicates that timed testing can yield reasonably effective item calibration with less than 100 subjects. And it indicates that there are dif-

Figure 6
Timed Testing Easiness Parameters (c_j/a_j)
Plotted Against Large-Sample Easiness Parameters
for Progressive Matrices



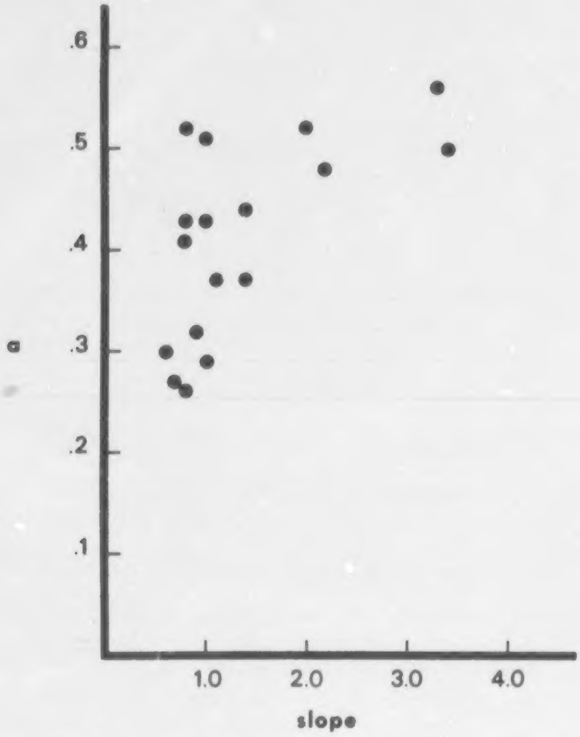
ferences in discrimination between items of this sort, which are reliable across changing sets of items, and between tests administered to junior high and college students.

Clocks

With only 17 usable items, the Clocks spatial test was the shortest of the three tests considered here and, probably as a result, in some ways the most unstable. Four of the items had estimated discrimination parameters greater than 2.0, and three other items had slopes very near 0, indicating that the estimation procedure took a short test and made it shorter by allowing a fraction of the items to dominate the scoring. It seems that the 2-parameter logistic model only remains "democratic" (uses all of the items in scoring) as long as there is a sufficiently large "silent majority" of test items to prevent the scoring procedure from freely reordering the examinees until a few items have near-perfect discrimination and the rest are omitted.

To some extent, this has happened here; but the estimation procedure stopped short of infinite discriminations. The high discriminations did result in some outliers among the estimated values for θ; four of the 78 estimates were between 2.4 and 4.4. However, the high-scoring students did respond correctly

Figure 7
Timed Testing Discrimination Parameters (a)
Plotted Against Large-Sample Slopes
for Progressive Matrices



to all of the items but one with near-zero discrimination, and frequently responded quite quickly, indicating that they were indeed individuals of very high spatial ability. So the seemingly extreme values may not represent serious difficulties.

As measured by the χ^2 goodness-of-fit statistics on the item correct-response × ability group contingency tables, there were certainly no difficulties. None of the item χ^2's were significant, and the total of 90.9 with 68 df was not highly significant, either ($p = .05$). The probability of correct response as a function of effective ability seemed to be approximated fairly well by the timed testing model with its estimated parameters.

The distribution of residuals from the log latency model showed approximately the same degree of non-normality as was the case for the Progressive Matrices; the χ^2 was 39.4 with 7 df ($p < .01$), which was highly significant. The pattern of non-normality was the same as it was for the Matrices data; most of the problem was caused by the combined effects of three-fourths of the responses being correct and the correct responses being faster. Again, the non-normality was not due to extreme outliers; so the parameter estimation, although probably not quite optimal, should not be seriously affected. The estimated value for the regression parameter b for the Clocks spatial test was only .025 (estimated SE = .003); while that implies that there was a significant relationship between response time and the probability of a correct response on the items of this test, the relationship was much smaller than for the other two tests. It seems that there were individual differences in the ability to respond correctly to these items and individual differences in speed of response; but, in striking contrast to the Matrices test, those two variables were unrelated. Indeed, the correlation between θ and s for the clocks was -.03. (Estimated correlations among all of the individual parameters are given in Table 2.)

Given the low correlation between θ and s, it would seem that only one of those variables could be the classical spatial trait. In this case, that was θ, on which there was a significant sex difference of the same magnitude and form usually found on speeded spatial tests ($F(1,76) = 4.65$, $p > .05$). Even allowing for some moderate non-normality, the males scored substantially higher. There was no hint of a sex difference on s with the Clocks. For this test, with three-dimensional rotations, it seems that the classical spatial trait simply determined whether the problem could be solved or not; using more or less time seemed to make little or no difference. This is in marked contrast to the results obtained earlier (Thissen, 1976a) with another (simpler, two-dimensional) spatial test in which both males and females had the same mean θ and in which the sex difference (and presumably the spatial trait) was in slowness controlling for effective ability. The Clocks may be a good paper-and-pencil spatial test because performance on these items is essentially unrelated to slowness or carefulness.

If effective ability and easiness are the spatial trait, what are the slowness parameters doing? To answer that question, the items, whose properties are fairly easy to define, may be examined. The easiness of an item seems inversely related to the amount of rotation it requires, a result reminiscient of the Shepard and Metzler (1971) result. Holding amount of rotation constant, the item slowness parameter is, in part, an increasing function of the number of symbolic "instructions" the item uses to achieve that rotation. The items in the test define the rotation to be applied by arrows on the surfaces of 1, 2, 3, or 4 spheres. A rotation of 180 degrees may be defined by any of those numbers of spheres; and in this subset of items, it was. The more spheres the item used to give the instruction, the longer it took to encode (presumably) and the longer it took for the examinee to respond. Regardless of the number of spheres involved, however, the 180 degree items had about equal easiness. Again, a computerized adaptive testing system using latency to estimate ability would have to recognize that for the Clocks spatial test, more "instruction spheres" do not necessarily make the item harder (that depends on the amount of rotation), but they do make the response slower.

Relationships Among the Tests

Correlations among parameters. The Progressive Matrices have been variously labeled as a test of abstract reasoning ability, as g, and as a number of other constructs. There is evidence in the present data (see Table 2) that when given no time limit, the Matrices become a test of slowness. Effective ability and slowness on the Matrices were correlated .94 in these data; the more slowly a subject responded, the more likely he/she was to get the item correct, and vice versa. Effective ability and slowness were not nearly so closely related for the Verbal Analogies (r = .67), and they were unrelated on the Clocks. One possible explanation is that some kinds of test items are affected more by slowness (and, possibly, carefulness) on the part of the examinees than are others. This speculation has some support in these data in the form of the relationships, shown in Table 2, between effective ability and slowness on the Verbal Analogies and the parameters of the names on the Matrices.

Table 2
Estimated Correlations Among the Person Parameters
for the Three Tests

Test and Parameter	Analogies		Matrices		Clocks	
	θ	s	θ	s	θ	s
Analogies						
θ	1.00					
s	.68	1.00				
Matrices						
θ	.54	.69	1.00			
s	.39	.68	.94	1.00		
Clocks						
θ	.42	.28	.39	.29	1.00	
s	-.09	.40	.36	.55	-.03	1.00

Both θ and s for the Progressive Matrices may be predicted quite well from effective ability and slowness on the Verbal Analogies; the multiple R's were .70 and .69, respectively. The standardized regression coefficients are given in Table 3. Effective ability on the Progressive Matrices was predicted by a small positive weight for effective ability on the Analogies and a larger positive weight for slowness on the Analogies. Individuals who answered items slowly on the Analogies responded correctly on the Matrices. Slowness on the Matrices was similarly predicted by slowness on the Analogies, with a small negative weight for Analogies effective ability: individuals who answered items entirely too slowly on the Analogies, given their ability to respond correctly, answered very slowly on the Matrices.

It could be concluded that the three tests in this particular set represent three types of timed tests. On the Progressive Matrices, working slowly and carefully was strongly related to the probability of responding correctly, and what is measured is largely slowness. On the Verbal Analogies, working slowly and carefully was related to responding correctly, but not so strongly; so two distinct traits are measured--analogical reasoning ability and slowness. On the Clocks slowness did not seem to affect the probability of a correct response; therefore, spatial ability and slowness are essentially measured separately.

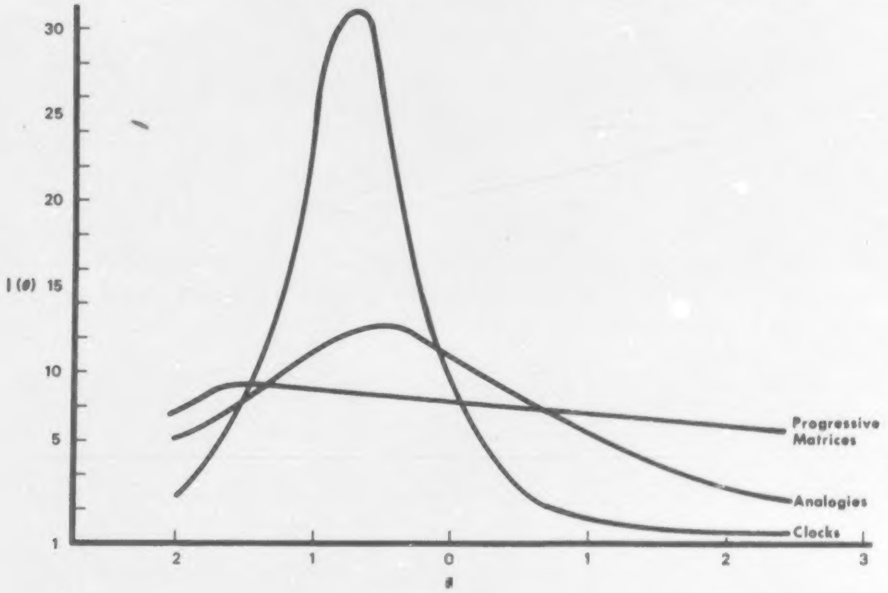
Information. The differences among the three types of tests are graphically portrayed in the test information functions for θ for the three tests, shown in Figure 8. The latency data had almost no effect on the ability estimation in

Table 3
Standardized Regression Coefficients for Prediction of the Matrices Scores from the Analogies Scores

Independent Variable	Matrices θ	Matrices s
Analogies θ	.13	-.13
Analogies s	.60	.77

the Clocks; as a result, the test information function for the spatial test has the classical "peaked" form of 2-parameter logistic test information functions for tests not constructed to spread the items very well. All of the information about ability comes from the item responses, and that information is only substantial near the difficulty level of the items, in this case between θ = -1 and 0.

Figure 8
Test Information Functions for the Three Tests



The Progressive Matrices, on the other hand, have a flat, regression-like test information curve. When b is estimated to be quite high, as it was in the case of the Matrices data in this example, effective ability was essentially estimated as though it were an element in a linear regression equation predicting log latency. The log latencies are assumed to provide the same amount of information about effective ability regardless of the value of θ; therefore, the test information curve is nearly flat.

The Verbal Analogies test information function shows a situation in which information was being obtained both from the item response data (giving the curve its familiar peaked form) and the latencies. The information provided by the response times serves to raise the entire curve, so that there is some information available in the system for estimating the effective ability of individuals with relatively extreme trait values. The test information function for the Verbal Analogies represents the desired outcome for the timed testing model; the other examples make it clear that we are only beginning to learn what can happen in a timed testing situation.

Conclusions

The latent trait model for timed ability testing described here is neither perfect nor complete; it still requires extensive (and expensive) computation to estimate the parameters of the model for fairly small samples, and it needs an additional parameter to absorb the relatively consistent difference between the residuals from the log latency model for correct and incorrect responses. Improved starting values for the maximum likelihood algorithm would go a long way toward solving the first problem. And the existence of the second problem is interesting. It is not too surprising that averaging over people or items, correct answers take less time than incorrect ones, because correct answers come from able people answering easy items and incorrect responses come from less able people responding to harder items. But this timed testing model corrects for the ability of the individuals and the easiness of the items, and it is still true that correct answers are associated with shorter latencies. This could be a real indication of some processing difference between correct and incorrect responses; future research could define the difference.

However, even with these potential problems, the timed testing model is ready for use. These data, as a matter of fact, were not collected to test the timed testing model; they were collected in an investigation of cerebral laterality and its relationship to cognitive abilities. The timed testing model was used to score the test because it used the available data most efficiently. As computers do the testing and time the responses, that will probably be the case with increasing frequency.

REFERENCES

Audley, R. J. A stochastic model for individual choice behavior. Psychological Review, 1960, 67, 1-15.

Audley, R. J. Some observations of theories of choice reaction time. In S. Kornblum (Ed.), Attention and performance IV. New York: Academic Press, 1973.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 24-51.

Fletcher, R., & Reeves, C. M. Function minimization by conjugate gradients. *Computer Journal*, 1964, *7*, 149-54.

Furneaux, W. D. Intellectual abilities and problem solving behavior. In H. J. Eysenck (Ed.), *The handbook of abnormal psychology*. London: Pitman, 1961.

Guilford, J. P., & Zimmerman, W. F. *The Guilford-Zimmerman Aptitude Survey. VI. Spatial Visualization, Form B*. Beverly Hills, CA: Sheridan Supply Co., 1953.

Haberman, S. *A conditioned Newton-Raphson algorithm for function minimization*. Unpublished manuscript, 1970.

Kagan, J., Rosman, B. I., Day, D., Albert J., & Philips, W. Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs*, 1964, *78* (1,Whole No. 578).

Laming, D. R. J. *Information theory of choice reaction times*. New York: Academic Press, 1973.

Luce, R. D. Response latencies and probabilities. In K. J. Arrow, S. Karlow, & P. Suppes (Eds.), *Mathematical methods in the social sciences, 1959*. Stanford, CA: Stanford University Press, 1960.

Rasch, G. *Probabalistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institute, 1960.

Raven, J. *Progressive Matrices* (Rev. ed.). London: Lewis, 1956.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 1969, *34* (4, Pt.2, Monograph No. 17).

Samejima, F. Homogeneous case of the continuous response model. *Psychometrika*, 1973, *38*, 203-219.

Shepard, R. N., & Metzler, J. Mental rotation of three-dimensional objects. *Science*, 1971, *171*, 701-703.

Sternberg, R. J. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum, 1977.

Thissen, D. *Incorporating item response latencies in latent trait estimation*. Unpublished doctoral dissertation, The University of Chicago, 1976. (a)

Thissen, D. Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, 1976, *13*, 201-214. (b)

Tinsley, H. E. A. An investigation of the Rasch simple logistic model for tests of intelligence or attainment. Unpublished doctoral dissertation, University of Minnesota, 1971.

White, P. O. Individual differences in speed, accuracy, and persistence. In H. J. Eysenck (Ed.), The measurement of intelligence. Lancaster, England: Medical and Technical Publishing Co., Ltd., 1973.

Whitely, S. E. Relationships in analogy items: A semantic component of a psychometric task. Educational and Psychological Measurement, 1977, 37, 725-739.

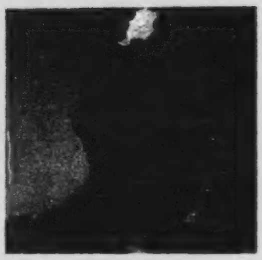
Whitely, S. E. A multi-component model for information processing abilities. Paper presented at the 1979 annual meeting of the Psychometric Society, Monterey, CA, April 1979.

Wiley, D. E. Latent partition analysis. Psychometrika, 1967, 32, 183-193.

ACKNOWLEDGMENTS

This research was supported in part by funds from the University of Kansas General Research Fund. I am grateful to R. Darrell Bock for providing the "Clocks" test items and to Susan Whitely for the Verbal Analogies, as well as for suggestions for their analysis. Thanks to Leslie Webster and Paul Isenberg for keeping the data collection going smoothly regardless of obstacles, and special thanks to Laura Baker for her technical asistance in the preparation of this paper and her comments on its contents.

DISCUSSION: SESSION 6



JOHN B. CARROLL
UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL

Both the Thissen and the Tatsuoka papers are excellent, presenting interesting and useful approaches to a problem with which I have long been concerned--the role of speed in testing ability and achievement. I can best comment on them by using them as stimuli for some general remarks about speed-ability relationships.

First, however, I will consider one technical issue that is touched upon in both papers--namely, the distribution of item response times over items and/or over individuals. Thissen has stated that "data analytic considerations suggest that response times are frequently approximately lognormally distributed," and his analyses consequently utilize logarithms of these times. Tatsuoka, on the other hand, has provided evidence that response times follow a Weibull distribution and has offered a tentative rationale for the appropriateness of such a distribution.

Before either the lognormal or the Weibull distribution for response times is accepted, consideration should be given to the basic metric for these times. It has appeared more reasonable to express response latency in terms of *performance per unit of time* (e.g., Landahl, 1940; Wainer, 1977). This is the conventional way of measuring, for example, the speed of a vehicle (miles or kilometers per hour), and it can also be applied to rate of work in performing test items. This metric has a theoretical zero point expressing a state of no motion, in the case of a vehicle, or a state of no performance at all, in the case of work on a test. When rate of performance is expressed in this way, I have generally found that individual differences tend to be normally distributed. For this reason, it has been my practice to take the reciprocals of response latencies and to use these in my data analyses, rather than the raw response times or even their logarithms. This has been done, for example, for picture-naming latencies (Carroll & White, 1973a, 1973b), interpreting the reciprocals as number of pictures that could be named per unit of time. In reporting central tendencies for response times, I use the harmonic mean, which is, of course, a back-translation from the arithmetic mean of the reciprocals of the response times.

From these considerations, I would suggest that Thissen (or anyone following his lead) might try to substitute the reciprocal transformation for the logarithmic transformation of response time. Possibly this would yield better fits to data, and in any case it would have a somewhat better theoretical underpinning.

On the other hand, I am much attracted to the possibilities that seem to be offered by the Weibull distribution as investigated by Tatsuoka. She has gone far in offering a reasonable rationale for the application of this distribution to time score data in suggesting that the test item is seen as a "system" that the individual tries to break down. Somehow, I have always thought of the matter in reverse--that is, the examinee is the system that the test item is trying to break down. The Rasch model, incidentally, presents the idea that whichever will break down first--the individual or the test item--is given by the relation between the person parameter and the test item parameter. In dealing with response times, however, the assumptions of the Weibull distribution make it reasonable to consider the relation unidirectional, in the sense that one "waits" for the examinee to "break" the test item. The parameter $\underline{c}$ in Weibull distribution theory appears to be of particular interest, for it specifies whether the individual's probability of mastering the test item increases, remains constant, or decreases with increasing time. As Tatsuoka noted, "it is intuitively plausible that items of all three kinds may exist in practice, depending on the difficulty and other properties of the item." Even mixed cases are possible, for example, one in which the probability first increases, then decreases, or one in which the probability has a constant low value and then increases rapidly, as for a "sudden insight" problem solution. Possibly the Weibull distribution could be applied to such cases by assuming a two-stage process, with different parameters for each stage. Probably, however, the estimation of separate parameters for the stages would be a formidable problem.

The theme of speed-ability relationships is an old, but relatively neglected, problem in psychometrics. It is not difficult to find investigations of it in the early literature (e.g., Dubois, 1932; McFarland, 1930). Thurstone (1937) formulated a psychometric model of ability, motivation, and speed involving a three-dimensional psychometric surface in which these variables could vary independently. Baxter (1941) pointed out that time-limit and work-limit scores have an artifactual part-whole relation that accounts for their high intercorrelation in most circumstances and discovered that sheer rate of work or performance on the Otis group intelligence test had a correlation of essentially zero with level of ability as determined from work-limit scores on the test. Davidson and Carroll (1945) pursued this matter further and confirmed these relations in the case of subtests of the Army Alpha. It was shown that several different speed and level factors could be identified in the subtests of this battery when they were administered in such a way as to obtain not only the conventional time-limit scores but also scores measuring rate of work and level of ability. The time-limit scores were shown to have factor loadings on both speed and level factors.

Since the Davidson and Carroll study, this line of investigation has been pursued only very infrequently in the factor analytic literature. One exception is the study by Lord (1956), disclosing separate dimensions of ability and speed in a number of domains of cognitive performance. One of Lord's findings, for example, was that ability level in spatial tests is a dimension separate from speed in performing those tests. Recently, Egan (1976), working at the item response level, found a similar two-dimensional structure for spatial ability tests.

Baxter had proposed that speed, power, and level scores for ability tests should be carefully distinguished; he defined a speed score as a sheer rate-of-work score without regard to the correctness of response. Conventional time-limit scores were to be designated as power scores; work-limit scores (number correct in unlimited time) were to be regarded as level of ability scores. Unfortunately, Baxter's proposals were never generally accepted in the psychometric literature; most often, time-limit scores are called speed scores and work-limit scores are called power scores. The more unfortunate error, however, is to continue to assume that conventional time-limit scores are appropriate measures of level of ability, when actually they usually reflect rate of work to a degree that is a function of the speededness of the test, or more properly, the time limit.

The confusion between ability and speed has arisen primarily because of the conventional methods of administering tests with time limits that are often set rather arbitrarily but, in any case, as short as is deemed feasible. Computerized testing, whether it is adaptive or not, offers a means of avoiding the problem of speeded testing because it can control the test items offered to the examinee and can measure the time taken to respond to them. What is most intriguing about Thissen's paper is its proposed methodology for extracting both ability and speed information from computer-administered tests. This methodology seems to be highly promising. I would be particularly interested if it could be developed to permit multidimensional results in either ability or speed domains or in both.

Although Thissen's remarks about the interpretation of the ability and speed parameters of the three tests that he analyzed were of interest, his data treatment might well have included a factor analysis of his table of intercorrelations. I have taken the liberty of performing such a factor analysis, even though the iteration for communalities had to stop at 10 iterations to avoid having at least one of the communalities exceed unity. A Varimax-rotated solution of the common factor matrix arrived at after 10 iterations is shown in Table 1. Obviously, the two uncorrelated factors, together accounting for about 77% of the total variance, may be interpreted as ability and speed, respectively. What is of particular note is that the factors generalize over different tasks. The best "pure" measure of ability is the parameter θ for the Analogies test, while the purest measure of speed is the s ("slowness") parameter for the Clocks test. Nevertheless, the factors show up in interesting ways on other tasks.

Especially noteworthy were the results for the Raven Progressive Matrices Test. There is probably more confusion and conflicting evidence in the literature on the Raven test than on any other commonly used test. Many authors (e.g., Jensen, 1978) regard the Raven test as one of the best measures of g, or general intelligence. Thissen's results, however, indicate that the test is factorially complex, measuring ability and speed in both the θ and s parameters. Thissen speculated that when the Raven test is given under no time pressure, it is primarily a measure of speed or its opposite, slowness, or perhaps carefulness.

Table 1
Varimax Rotated Principal Factor Analysis
of Thissen's Table 2 of Correlations among
Estimated Person Parameters for Three Tests

Test	Factor Loadings Factor 1 (Ability)	 Factor II (Speed)
Analogies		
θ	.97	.01
s	.62	.52
Matrices		
θ	.59	.71
s	.41	.91
Clocks		
θ	.45	.09
s	-.09	.67

Note. The preliminary principal component analysis of the correlation matrix with unities in the diagonal yielded eigenvalues of 3.35, 1.30, 0.70, 0.45, 0.16, and 0.03. Iteration to two factors for communalities stopped iteration 10 before one of the variables would have attained a communality greater than 1.0. This matrix is presented as sufficient to exhibit the overall pattern of the results.

Horn (1978) made an extensive factor-analytic study of speed, power, and carefulness (among other things), supported by the Army Research Institute. Two of his conclusions are the following:

"1. There is considerable cohesion among indicants of average speediness in providing response (either correct or incorrect) in tasks of nontrivial difficulty.

"2. Intellectual speediness indicated in this manner has only a very low, perhaps only chance, relationship to the goodness of intellectual performance that is indicated by the number of correct answers provided in a wide range of putative measures of intelligence."

Horn's study has elaborated the concept of speediness to a much greater extent than can be reviewed here; he also considered the role of strategies that examinees may adopt in attacking items and the dependence of these strategies on the character and content of the items or tasks. Actually, his evidence suggests the existence of two speediness factors--CDS (correct decision speed) and QDS (quit decision speed)--the latter pertaining to situations in which the individual decides to give up in a problem-solving task.

This result gives me added confidence in suggesting that the methods developed by Tatsuoka and by Thissen might be applied in further study of ability, speed, and carefulness factors at the item level. Using a mixed model Weibull distribution with different _c_ parameters, the CDS and the QDS factors might be more reliably differentiated. One characteristic of items that might be relevant here is whether the items are "self-revelatory," that is, whether they have a solution that can be recognized as correct by examinees once they discover it, in contrast to the usual test item for which the correct answer is not obvious to examinees unless they have the required level of ability or knowledge. It would be profitable to apply Thissen's methods to a wide range of ability tests in order to determine the dimensions of ability and speed in such tests. I would expect the structure to be much more multidimensional than Thissen's preliminary results show.

Most applications of latent trait theory thus far are limited to the unidimensional case, or at least to cases in which unidimensionality is assumed. There is abundant evidence that abilities in the cognitive domain are multidimensional. It is my hope that work in latent trait theory in the future can address the multidimensional case more fully.

REFERENCES

Baxter, B. J. An experimental analysis of the contributions of speed and level in an intelligence test. _Journal of Educational Psychology_, 1941, _32_, 285-296.

Carroll, J. B., & White, M. N. Word frequency and age of acquisition as determiners of picture-naming latency. _Quarterly Journal of Experimental Psychology_, 1973, _25_, 85-95. (a)

Carroll, J. B., & White, M. N. Age-of-acquisition norms for 220 picturable nouns. _Journal of Verbal Learning and Verbal Behavior_, 1973, _12_, 563-576. (b)

Davidson, W. M., & Carroll, J. B. Speed and level components in time-limit scores: A factor analysis. _Educational and Psychological Measurement_, 1945, _5_, 411-427.

DuBois, P. H. A speed factor in mental tests. _Archives of Psychology_, 1932, _22_ (Whole No. 141).

Egan, D. E. _Accuracy and latency scores as measure of spatial information processing_ (Research Report No. 12224). Pensacola, FLA: Naval Aerospace Medical Research Laboratory, February 1976.

Horn, J. L. _Final report on a study of speed, power, carefulness, and short-term learning components of intelligence and changes in these components in adulthood_. Denver, CO: University of Denver, 1978.

Jensen, A. R. *g*: Outmoded theory or unconquered frontier? *Creative Science and Technology*, 1978, *2*, 16-29.

Landahl, H. D. Time scores and factor analysis. *Psychometrika*, 1940, *5*, 67-74.

Lord, F. M. A study of speed factors in tests and academic grades. *Psychometrika*, 1956, *21*, 31-50.

McFarland, R. A. An experimental study of the relation between speed and mental ability. *Journal of General Psychology*, 1930, *3*, 67-97.

Thurstone, L. L. Ability, motivation, and speed. *Psychometrika*, 1937, *2*, 249-254.

Wainer, H. Speed vs. reaction time as a measure of cognitive performance. *Memory & Cognition*, 1977, *5*, 278-280.

Session 7:

Person-Item Interaction

Using the Rasch Model to Identifying Person-Based Measurement Disturbances

Ronald J. Mead
Minneapolis

Robust Estimation in the Rasch Model

Howard Wainer
Bureau of Social Science Research
Benjamin D. Wright
The University of Chicago

Appropriateness Measurement: Basic Principles and Validating Studies

Michael Levine
University of Illinois
Fritz Drasgow
Yale University

Discussion

James Lumsden
University of Western Australia

USING THE RASCH MODEL TO IDENTIFY PERSON-BASED MEASUREMENT DISTURBANCES

RONALD J. MEAD
MINNEAPOLIS

There is currently in psychometrics a controversy due to a fundamental difference of opinion about what measurement is and how it is to be used. George Bernard Shaw (1903) addressed an important aspect of this controversy in "Maxims for Revolutionists" when he wrote, "The reasonable man tries to adapt himself to the world; the unreasonable man persists in trying to adapt to the world to himself ...."

In psychometrics the "reasonable" approach involves constructing a model sufficiently complex to explain any data that might be produced by a group of people taking a set of test items. The "unreasonable" approach fixes on a particular model and struggles to make data conform to it, the choice of the model being determined by philosophical rather than empirical considerations. To avoid any confusion about where Shaw stood with respect to reasonableness, he went on to say, ". . . therefore, all progress depends on the unreasonable man."

The simple logistic model can be viewed from either position. For the "reasonable" psychometrician, it is known as the 1-parameter model, which is a special case of the 2- (or more) parameter model; and it should be considered when it fits the data, if for no other reasons than economy and parsimony. This viewpoint has a respectable ancestry in model building (with linear models), where the magnitude of the unexplained error is the criterion for deciding if parameters should be added or deleted.

For the "unreasonable" psychometrician, the simple logistic model is known as the Rasch model. It is the very definition of measurement; hence, measurement is not possible when data do not fit it. When viewed from this aspect, it is a very special model indeed, but is not a special case of anything.

This difference in philosophy might be compared to the difference between stepwise multiple regression and experimental design, both of which have been very useful to the social sciences. The first is concerned with fitting data using whatever model is necessary; the second, with organizing the situation to obtain data that conform to a model which will make the estimation and the inferences as easy and as unambiguous as possible.

In analysis of variance terms, the Rasch model is a main class model with two fixed classes and no interactions. All the data needed to estimate the effects are contained in the marginal sums; and since all the marginal information is needed in estimating the main class effects, any additional parameters, regardless of how they are subscripted, must represent interactions between the person and the items. This raises the old analysis-of-variance dilemma of how to interpret main effects in the presence of an interaction.

Returning to psychometrics, if the model contains any additional parameters, such as item discrimination, person sensitivity, random guessing by some persons, or nonzero asymptotes for some item characteristic curves, then either the item characteristic curves (ICCs) or the person characteristic curves (PCCs) or both will not be parallel (after linearization). Hence, the comparison of the abilities of two persons will depend on the particular items used as the basis for the comparison. This is a statement of what is meant by failing to achieve "specific objectivity" as defined by Rasch (1960), but it is also a statement of an interaction as used in analysis of variance.

As with interactions, these additional parameters have meaning only when at least one of the ways of classification (e.g., persons) can be considered a random sampling from some relevant population. Inferences about ability are therefore normative in the sense that they pertain only to comparisons within that population.

The most fundamental distinction between the Rasch model and the other psychometric models is that the Rasch model concentrates on the person, whereas the other approaches deal with groups of people. Rasch (1960) quoted two psychologists on this topic. Citing Skinner (1956), he stated, "Any order to be found in human and animal behavior should be extracted from investigations into individuals, and (current) psychometric methods are inadequate for such purposes, since they deal with groups of individuals." He quoted Zurbin (1956) as saying, "Recourse must be had to individual statistics, treating each patient as a separate universe. Unfortunately, present-day statistical methods are entirely group-centered, so that there is a real need for developing individual-centered statistics." This is no less important in education. When the intent is to describe the progress or achievement of one student, it should not matter to what populations he/she is assigned.

In solving this problem, Rasch formulated some general principles of comparison, which can be rephrased as follows:

1. For any relevant item, a more able person always has a better chance of success than a less able person, and
2. Any person has a better chance of success on an easier item than on a more difficult one.

These statements indicate nothing about the age, sex, race, or religion of the person, only that the items be appropriate to him/her for the variable.

Although these conditions may seem so obvious and necessary as to be almost trivial, the family of models proposed by Rasch are the only ones that meet them. All other models lack these properties, which Rasch has called "specific objectivity"--objective because the comparison of any two people is independent of the items used and specific because the items must be appropriate for making one particular comparison.

There are, as Rasch (1960) and Fischer (1976) have shown, a large family of exponential models that have specific objectivity with a variety of types of observations. The only one to be discussed in this paper is the model for dichotomously scored items, which looks like the Birnbaum model and, as mentioned previously, is commonly known as the Rasch model. Because it does have the potential of achieving objectivity, however, it is not a special case of the Birnbaum model, and choosing between them should be on the basis of whether or not objectivity is of value.

Person-Based Disturbances

There are two major concerns motivating an interest in the analysis of person fit. First, the simplicity of the Rasch model makes it very demanding on the data. Although some very desirable measurement properties are associated with the model, they are attained only if the data fit. A thorough search for misfit is therefore essential.

Second, since measurement of the person is the objective, it is only prudent that before any decisions are made about the person (based on a series of simple responses to artificial situations), it be verified that the responses mean what is intended. Regardless of how well or how often the items have been used in the past, there is no guarantee that they operated as planned with a particular person on a specific occasion.

It is preferable to use the Rasch model for this purpose precisely because of its simplicity and because of the logical relationship between its demands and its benefits. If the demands are met, the benefits necessarily follow. Since it is simple, it will not appear to explain data that were generated by a multidimensional process. When it fits, we know exactly what we have. When it does not fit, at least we know what has been taken out; therefore, all the information about misfit should be left in the residuals from the model.

In order to use the model to understand what could have gone wrong when the data do not fit, some consideration must be given to how various forms of misbehavior might appear in the data. It must be remembered that when the data do not fit, the same logical position does not exist. Although it can be predicted how the data will look if certain things happen, it does not follow that if the data look that way, these things have necessarily happened.

Three much discussed disturbances will be considered in this paper: random guessing, speededness, and bias. These are relatively general, since many other problems can be stated analogously and all can be handled by a single strategy. In addition, it will be suggested how the fitted ICCs might be affected if a substantial proportion of the sample were engaging in these activities.

Random Guessing

This can only occur (1) when the person has no knowledge that would help him/her choose a response or (2) when he/she does not read the question before responding. Not everyone will guess randomly, and among those who will, the propensity to do so is not necessarily the same. One reasonable description of this process is as follows: If the person has a reasonable amount of knowledge about the item, he/she will respond according to the model (solid ogive in Figure 1a). At some level of difficulty, the person will decide that he/she knows nothing about the item and will respond randomly, with the probability of success (dashed line) being determined by the number of alternatives. The point (G_v on the figure) at which this change-over occurs surely varies from person to person according to confidence in ability and tendency toward risk-taking.

The residuals, after removing the effects of the item difficulties and the person's ability, are as shown in Figure 1b. The positive residuals for the difficult items imply a surprising degree of success on these items. Surprise increases in moving toward more and more difficult items, and the rate of success remains the same. The negative residuals correspond to responses that should fit the model, but the expectation was upset by the person's unwarranted successes on the difficult items. Dropping the difficult items from consideration and recomputing the person's ability should result in an acceptable measurement.

If a substantial proportion of the sample were behaving in this way, it would be expected that ICCs would be as shown in Figure 2. Estimates of item difficulty would be biased downward (that is, the item would be considered easier than it actually is). Heterogeneity in the item discriminations would also be observed, with the more difficult items appearing to have the lower discriminations. The extent of the disturbance in both would depend on the propensity of the sample to guess and on the proportion of the sample in a position where guessing was a viable strategy. The methods devised by Waller (1974, 1976) to correct for guessing should effectively eliminate both problems.

Speededness

It frequently happens with group-administered tests that not everyone has ample time to completely answer every item. Although time limits are normally chosen to minimize this, they invariably involve some compromise between administrative convenience and handicapping a few persons.

How a time limit affects a person undoubtedly depends on the individual. The person might simply rush through the test without spending enough time on any item. The residual response string would have the appearance of both random guessing and carelessness, with some difficult items answered correctly and some easy items answered incorrectly. The effect on the estimate of ability would be to underestimate it. Except for a general fuzziness in the quality of the measurement, this situation would be difficult to detect and diagnose psychometrically.

A slow methodical person might carefully consider each item before responding and consequently could leave several unanswered at the end. Detecting this does not require high-powered psychometrics. Skillful test-takers might complicate this picture by filling in as many answers as possible between the time the

Figure 1a
Person Characteristic Curve with Guessing

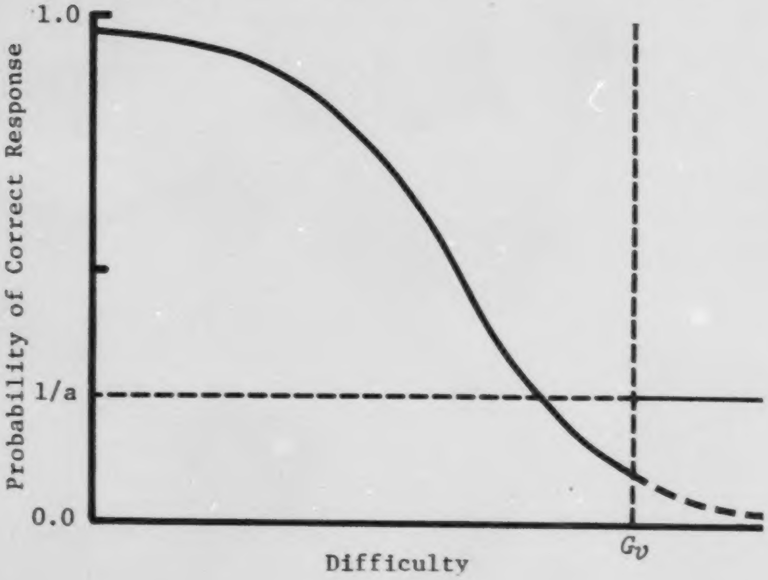


Figure 1b
Mean Residuals from Model for Person with Guessing

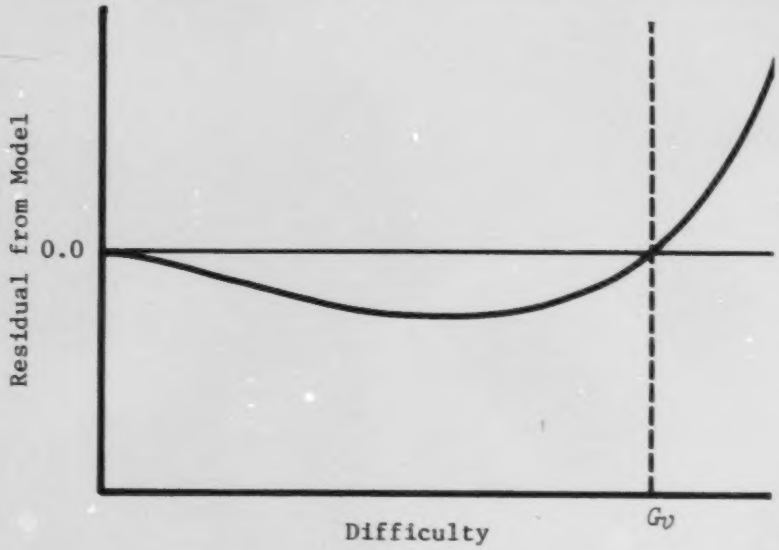
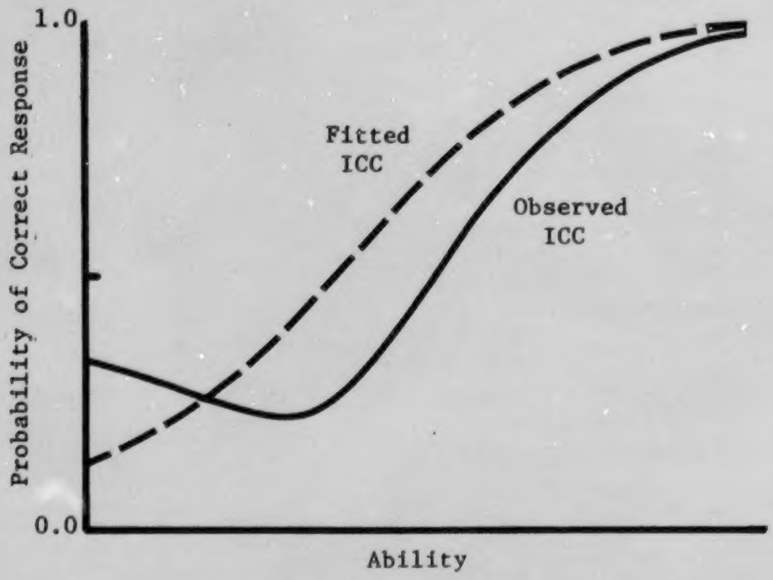


Figure 2
Typical Item Characteristic Curve with Guessing



proctor says stop and when the answer sheet is taken away. The response string should be acceptable up to the point at which the behavior changes. From there on it should have the appearance of random guessing.

Deciding what to do about either of these cases requires some consideration of the nature of the variable being sought. It is sometimes argued that the capability of performing the tasks rapidly is an important component; so all items presented to the person should be considered in the measurement. However, the ability to do something and the ability to do it quickly are not necessarily the same. They may both be important and are often highly correlated; but if they are different variables, they can not be combined into a single, valid measure and treated with a unidimensional model.

The effect of ignoring the speededness on the estimate of item statistics would be to make the items at the end of the test appear more difficult and more discriminating than they actually are. They are too difficult because too few people responded to them correctly, and the discriminations are too high because the persons who did respond correctly will tend to have more correct answers on the whole test (because they actually took a longer test).

Bias

An item is biased against a person if, for some reason, the person is at a disadvantage on that item relative to other items and other persons. This must involve an additional latent variable, and the effect of the person's position on this variable is to lower his/her chances of success on the affected items. One familiar example is a vocabulary test in which the word "sonata" is found to be biased, since not all cultural groups would have had equal exposure to it.

Most discussions of item bias (e.g., Draba, 1978) in connection with the Rasch model have suggested techniques such as the following:

1. Calibrate all items on each group separately;
2. Compare the resulting item difficulties;
3. Items which display significant shifts in difficulty between groups are considered potentially biased; and
4. If content experts concur, the items are revised or dropped.

Although this approach is very useful, it is not the complete answer. It has two obvious shortcomings. First, it depends on internal definition of the true variable. It can only work well if most items are "fair." If the bias is present uniformly in all items, the unfavored group will simply appear lower in ability. Second, this analysis is, again, population-based. It requires the definition of groups that must be arbitrary to some extent and then assumes that any bias in the items is the same for all members of the group. It seems preferable to treat every person as his/her own group and to require each item to demonstrate its validity for every person.

This can be accomplished by rearranging the steps and bringing the content experts in first. Their function would be to cluster the items according to some criterion so that the subsets would be homogeneous with respect to any suspected extraneous variables. This might be on the basis of vocabulary, as suggested earlier; or in the case of reading comprehension, it might be on the basis of the subject matter of the passages.

To take another example, if a mathematics reasoning test were comprised of word problems, the items might be grouped by their readability. In this case, the person obviously has two abilities and the item two difficulties. The person can be successfully measured on one of the variables only if his/her performance is not affected by the other. If the person is sufficiently skillful at reading, he/she will read and understand all problems, regardless of the problem's resistance to being read. Therefore, the person's performance will be determined by his/her reasoning ability and the difficulty of the problem. It would not matter if the items were equally difficult to read or not.

On the other hand, if the person were a very skillful problem solver and a poor reader, it would be a test of his/her reading ability. His/her performance would depend on whether or not he/she could read and understand the problem. If the calibrating sample were like this, the difficulties assigned to the items would be due to their locations on the reading variable.

The discriminations observed for the items could be influenced in either direction. One interesting situation would arise if the calibration sample were comprised of two groups, equally able in problem solving but substantially different in reading ability. Assuming that the first group had no difficulty reading any of the items but that the second had trouble with several, those items will have apparent discriminations. Since the groups are the same in reasoning, they will only be separated by the difficult reading items; and for purposes of this paper the high discriminations would be entirely spurious.

Multidimensionality

Clearly, the problem here is that the dimensionality of the latent space is greater than the dimensionality of the model. In practical applications it is more convenient to try to control the situation rather than to generalize the model. This conception of bias can, however, be generalized to include any strategy for forming subsets of the items. For example, difficult items could be considered biased against nonguessers; easy items, against careless test takers. Items at the end of a test are biased against slow workers; those at the beginning, against slow starters. Items that have never been used before are biased against examinees who belong to fraternities with good test files. There never seems to be any problem for people who are interested in a test to generate hypotheses about possible dimensions in it.

At this point a simple strategy can be described for checking each person's response string for multidimensionality. This person analysis is not a replacement for a thorough item analysis but, rather, is an addition to it. The hypotheses here about Person × Item subset interactions are distinct from the hypotheses in item analysis about Item × Person Group interactions. The power of the person analysis comes from replicating items that are alike in some sense; in item analysis it depends on replicating similar people.

The principle employed in the analysis is the objectivity of the measure. If it is truly objective, then all subsets of items should yield statistically equivalent estimates of the person's ability. If not, it would be concluded that the presence of multidimensionality is related to the manner in which the subsets were defined.

The most obvious method of doing the arithmetic would be to actually compute the ability associated with the person's score on each subtest and to perform an analysis of variance to test for between-subtest differences. This has the immediate drawback of being unable to deal with 0 or perfect scores. Gustafsson (1979) has recently proposed a set of procedures based on conditional maximum likelihood estimation, which has many desirable statistical properties. This, however, can be expensive, and it is still somewhat restrictive in the maximum number of items it can handle.

A convenient and economical analysis can be developed from the unconditional estimation equations. The basic equations needed are shown in Table 1. The notation used follows the conventions of Rasch where possible:

$\underline{v}$ designates the person,
$\underline{i}$ designates the item,
$\underline{X}_{\underline{vi}}$ is the score obtained by person $\underline{v}$ on item $\underline{i}$ (equals 0 if incorrect, 1 if correct),
$\underline{b}_{\underline{v}}$ is the ability of person $\underline{v}$ estimated from all items, and
$\underline{d}_{\underline{i}}$ is the estimated difficulty for item $\underline{i}$.

Table 1
Unconditional Person Analysis

Scaled Residual

$$Y_{vi} = (X_{vi} - P_{vi})/W_{vi}$$

where

$$P_{vi} = \exp(b_v - d_i)/(1.0 + \exp(b_v - d_i)) ,$$

and

$$W_{vi} = P_{vi}\ (1 - P_{vi})$$

Misfit Due to Subtest J

$$V_{vj} = \sum_{i \varepsilon j} y^2_{vi} W_{vi} / \sum_{i \varepsilon j} W_{vi}$$

Effect Due to Subtest J

$$Y_{vj} = \sum_{i \varepsilon j} y_{vi}\ W_{vi}\ /\ \sum_{i \varepsilon j} W_{vi}$$

Between-Subtest Mean Square

$$V_{vB} = \sum_{j} Y^2{}_{vj} \sum_{i \varepsilon j} W_{vi} / \sum_{i} W_{vi}$$

The scaled residual is simply the difference between the person's observed item score, X_{vi}, and his/her predicted item score, P_{vi}, predicted from his/her performance on the total test. The difference has been rescaled by multiplying by $1.0/(P_{vi} \times (1-P_{vi}))$, which is the derivative of b with respect to P, so that the residual is expressed in logits to a first-order approximation.

The misfit statistic has the form, but not the distribution, of a sum of squared z-statistic; that is, it is the sum of X minus P^2, divided by the sum of PQ. It will be large when the ability does not adequately explain the person's part in every X_{vi}.

The effect due to subtest j is simply the first adjustment that would be made if estimating the person's ability for subtest j were attempted using the total test ability as the starting value. This form is the Newton-Raphson solution to the unconditional maximum likelihood estimation equations. Since it is not iterated, there is no problem with zero or perfect scores. The between-subtest mean square asks if all these effects are null.

There are some problems with these statistics. In particular, neither the form of their null distribution nor the appropriate degrees of freedom is known.

Wright (1979) and Haberman (1979) have been investigating various weighting and standardizing schemes with the hope of bringing the distribution into line with a standard distribution. It seems important from their work that the numerators and denominators be summed separately, as they are in Table 1.

The degrees of freedom are a problem, since the usual analysis of variance type counting assumes every observation contains the same amount of information, which clearly does not apply here. One promising candidate seems to be to base a calculation of pseudo-degrees of freedom on the information function. In some cases this may be as simple as 4.0 × PQ.

Additional work is needed in this area. Two obvious and useful activities are the careful simulation of known situations over a broad range of conditions and a comparison of these statistics with those produced by the conditional approach of Gustafsson.

In practice, however, these studies are of secondary interest. Since it is well known that data do not fit the Rasch model anyway, it is of marginal utility to continue demonstrating this. What is useful is a general index of the quality of measurement for each person, that is, an indication of how close the data reflect objectivity. A weighted fit mean square based on the scaled residual in Table 1,

$$V_v = \frac{\Sigma(X_{vi} - P_{vi})^2}{\Sigma(P_{vi}\ (1 - P_{vi}))}\ , \qquad [1]$$

seems to accomplish this. Following that, some specific statistics are needed to help diagnose the problems when they occur. The unconditional between-set analysis shown in Table 1 has proved useful for this in a variety of applications. The additional research work needed includes the application of statistics like these to real data and the interpretation of the results to knowledgeable people to see if they make sense.

Examples of the Person Analysis

Certification Examination Data

The first example, shown in Table 2, is taken from an actual administration of a professional certification examination. The examination included about 1,000 items, some of which were omitted in scoring. The items were arranged in six booklets, intended to be as parallel as possible, and administered over two days. Six test committees, operating independently and representing six different content areas, actually wrote the items. Subject to printing considerations, the six areas were distributed evenly over the six booklets. The certification decision was based on the total raw score.

The only justification for considering this test to be unidimensional is empirical. After analyzing literally tens of thousands of examinees, less than 1% appear to be seriously flawed. This may be attributed to the homogeneity of the training program.

Table 2
Among and Within Item Subset Analyses For Example 1

```
NAME COUNT  SCORE ABILITY S.E  WTHIN BETWN

ADMINISTRATIVE BOOK 0.15  0.02 5.8   14.0                   +
BK 1  144.   29.   -1.54  0.23 5.8           --*--          +
BK 2  157.   96.    0.42  0.18 0.7                          +--*--
BK 3  149.   81.    0.45  0.18 0.3                          +--*--
BK 4  139.   82.    0.34  0.19 1.1                          +-*--
BK 5  149.   90.    0.59  0.19 1.6                          +  -*--
BK 6  130.   74.    0.49  0.20 0.3                          +--*--

NAME COUNT  SCORE ABILITY S.E. WTHIN BETWN

                    0.15  0.08 5.8   -1.3                   +
AAAA  150.   70.    0.09  0.18 1.0                        -*+-
BBBB  149.   79.    0.35  0.18 2.0                          +-*--
CCCC  146.   83.    0.00  0.19 1.8                       --*+
DDDD  139.   77.    0.06  0.19 2.1                       --*+-
EEEE  144.   73.    0.25  0.19 1.6                         -+*--
FFFF  140.   70.    0.10  0.19 5.7                        --*-

NAME COUNT  SCORE ABILITY S.E. WTHIN BETWN

ITEM TYPE           0.15  0.00 5.8    2.2                   +
 A    239.  129.    0.18  1.14 4.9                         -*--
 B    229.  128.    0.24  0.15 2.9                         -+*-
 C    140.   67.    0.01  0.19 1.3                       --*+-
 K    174.   80.    0.06  0.17 2.3                       --*+-
 N     30.    8.   -1.04  0.45 -0.2         ----*-----      +
 G     56.   40.    0.88  0.32 -0.2                         +    --*---

NAME COUNT  SCORE ABILITY S.E  WTHIN BETWN

NEW OR USED ITEMS   0.15  0.08 5.8   -0.1                   +
NEW  481.   238.    0.08  0.10 5.9                        -*+
USED 387.   214.    0.22  0.11 2.1                          +*-

NAME COUNT  SCORE ABILITY S.E  WTHIN BETWN

DIFFICULTY          0.15  0.08 5.8    4.7                   +
-2.0   48.   39.   -1.14  0.39 9.4          ----*----       +
-1.0  106.   80.   -0.27  0.22 2.0                    --*-- +
 0.0  237.  149.    0.08  0.14 0.4                        -*+-
 1.0  314.  136.    0.19  0.11 0.7                         -*--
 2.0  163.   48.    0.59  0.17 5.1                          +  -*--
```

The method that was used to decide which tests were flawed was both statistical and substantive. All interesting fit statistics were computed, their distributions examined, and the suspicious cases displayed for discussion by a task force selected for that purpose. Beginning with the largest misfit and working down, the task force examined each case in order until it was satisfied that the statistics were best explained as minor random fluctuations. In general, this occurred at about three standard deviations above the mean. The statistics were presented as pseudo-*t* statistics rather than mean squares, to keep the standards as consistent as possible. In all cases the means were near 0 and the standard deviations about 2.

Table 2 is a portion of the display for the person with the largest between-subtest statistic (14.0) of any in this administration. Each panel in this display represents a different criterion for defining subtests (i.e., test booklet, item type). In all cases the column labeled "ABILITY" presents, first, the estimate of the person's ability based on the total test (.15), followed by the ability based on each subset. The fit statistics are in the columns "WTHIN" (within) and "BETWN" (between). The total fit statistic is the first number in the WTHIN column and the between-subtest statistic is in the BETWN column. The remaining numbers in the WTHIN column are the fit statistics within each subset. They are analogous to total only if that subtest were being considered.

The explanation for the large between-booklet statistic is simple. A separate answer sheet was used for each booklet. The first one for this person was torn slightly, causing the scanner to misread the form identification, so the result was an essentially random score for that booklet. Rescoring this sheet correctly eliminated all disturbances in this record.

Table 3
Fit Statistics for Example 2

Ability	-1.09
SE	0.37
Total Fit	0.0
Sequence	
Between Subsets	2.4
Within Subsets	1.1
Difficulty	
Between Subsets	-5.9
Within Subsets	0.2

Mathematics Placement Test Data

For those who can not depend on having a 1,000-item test, Table 3 and Figure 3 give a different type of example. This was from a mathematics placement test for beginning college freshman. It consisted of 40 items in two separately timed segments. The reward for doing well was placement in a more difficult mathematics course.

Figure 3
Item Residuals in Sequence Order for Example 2

```
 5.0 +
     |
     |
     |
     |
 2.5 +
     |                               C
     | AA                           C
     |         A              C
     |   A    A              C   C
 0.0 +----------------------C---------------------------
     |A     A    AB BBBBBBBB   CC CC  D DDDDDDDD
     |     A    A  B                   D
     |
     |
-2.5 +
     |
     |
     |
     |
-5.0 +-----------+-----------+-----------+-----------+--
      1          11          21          31          41
```

The fit statistics were again pseudo-*t*'s, but the standardization was done differently so that although the observed mean was again near 0, the standard deviation was about .6. The item subsets were defined by sequence and by difficulty, with four subsets in each. The fit between subsets defined by sequence and difficulty were both large enough to be interesting. The plot by sequence explains the problem: The person took only the first half of each segment. These statistics do not indicate if the problem was due to lack of ability, lack of interest, or the inability to read English; and the decision regarding appropriate action on the part of the test user requires answers to these questions.

The difficulty fit statistic was large negatively because the person appeared too "sensitive" to item difficulty. The items were arranged in roughly increasing order of difficulty in each segment. Therefore, by only attempting half the test, the person tended to answer the easier items correctly and to answer all the more difficult ones incorrectly.

Conclusions

There is, in practice, an important distinction between what might be called statistical dimensions and conceptual dimensions. Statistical dimensions are any that can be found in the data; conceptual dimensions are everything that can be thought of. Successful measurement requires the former, but safety requires constantly checking to be sure the latter exists only in our heads.

For example, in a homogeneous situation it might be possible to mix together verbal items and quantitative items. If the same results are obtained regardless of the proportion of each, the result could be called "verbal ability," "quantitative ability," or "general intelligence." With every test administration, however, it would be necessary to prove that it still does not matter--that whatever causes these things that seem different to look the same is still operating.

Consideration of how this can be done leads to the observation that statistical unidimensionality is necessary and sufficient for fit to the Rasch model. The necessity need not be argued here. With respect to sufficiency, however, if the data are unidimensional, a model of the Rasch family will fit it.

To support this rather remarkable contention, two arguments can be made. The first is that the Rasch model is the only latent trait model for which it is possible to uniquely and unambiguously rank-order all the objects and agents along a single continuum. With this ordering, unidimensionality is obvious. Without it, phrases such as "more" or "less of the variable" do not make much sense, since there will be situations in which people do better on a more difficult item than on an easier one. The ability to rank individuals and items when a single variable is involved and the inability to generalize this concept to more than one dimension seems a critical point in this discussion.

The second argument concerns item discrimination. The first one may have been objected to on the grounds that variation in item discrimination does not make the data multidimensional but does make it not fit the Rasch model. Throughout this paper the point has been made that many extraneous variables will cause apparent heterogeneity; but that is the reason for hesitation in using item discriminations, not why they should not be used.

Items with extreme observed discriminations can always be explained in terms of additional variables. Usually, such items are obviously flawed: Some irrelevant and perhaps nonreproducible aspect of the item has interacted with special characteristics of the sample. Occasionally, these items provide a useful and constructive insight into the variable. Generating new items to take advantage of this new knowledge may very well lead to a refined (i.e., changed) definition of the variable that serves our purposes better than the old. An important point here is the ability to abstract whatever it is that distinguishes this item and to use it in the new items. It would then be expected that the new items would have discriminations that are similar to one another.

This paper has attempted to make two statements about dimensionality and the Rasch model. First, an explanation has been suggested of why such a simple model has worked as well as it has in many complex situations employing such artificial agents as multiple-choice items. Second, the unique relationship between the model and unidimensionality has been described. Whether or not this represents a fundamental and universal truth, it would be extremely productive to adopt this as the definition of unidimensionality. It would probably lead to better tests than exist now.

In adaptive testing the theory and simulations suggest that adequate measurements can be obtained with a handful of items. Attempts to do this in practice have not worked very well. This suggests that the technical knowledge about how to do this seems to be growing faster than the substantive wisdom about why it is desirable to do it. The methodology is now available to converge very rapidly on the person's location on a unidimensional trait; however, there do not seem to be very many unidimensional traits.

If it is intended that measuring people and making decisions about them based on those measurements continues, then a serious analysis of the quality of the measure for every person should be routine. Rather than being satisfied if the person is given a very short test, the power of adaptive testing would be better employed to explore the dimensionality of the space for the person. Is the first result reproducible over a useful range of the variable and for another selection of items? For most persons, the results will simply reassure the psychometrician that everything is in order; for other persons, something interesting may be learned about them and about the variable being pursued.

REFERENCES

Draba, R. E. The Rasch model and legal criteria of a "reasonable" classification (Doctoral Dissertation, University of Chicago, 1978). *Dissertation Abstracts International*, 1978, *39* (1-A), 245.

Fischer, G. H. Some probalistic models for measuring change. In D. DeGruijter & L.I. van der Kamp (Eds.), *Advances in psychological and educational measurement*. London: Wiley, 1976.

Gustafsson, J. E. *Testing and obtaining fit of data to the Rasch model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Haberman, S. Personal communication, June 1979.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institute, 1960.

Shaw, G. B. Maxims for revolutionists. *Man and superman*. New York: Dodd, Mead, & Co., 1903.

Skinner, B. F. A case history in scientific method. *American Psychologist*, 1956, *11*, 221-233.

Waller, M. I. *Removing the effects of random guessing from latent trait ability estimates* (ETS RB-74-32). Princeton, NJ: Educational Testing Service, 1974.

Waller, M. I. *Estimating parameters in the Rasch model: Removing the effects of random guessing* (ETS RB-76-8). Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1968.

Wright, B. D. Personal communication, June 1979.

Zurbin, J. *Experimental abnormal psychology*. New York: Columbia University Press, 1956.

ROBUST ESTIMATION IN THE RASCH MODEL

HOWARD WAINER
BUREAU OF SOCIAL SCIENCE RESEARCH

BENJAMIN D. WRIGHT
THE UNIVERSITY OF CHICAGO

Latent trait models as a class, and the Rasch model in particular, have begun to have substantial impact on the construction and scoring of mental tests. Through the use of latent trait models, measures of individual ability as well as item difficulty that have important practical and statistical properties can be obtained. For example, if the Rasch model fits, the measures of ability and difficulty obtained are interval-scaled, thus making the quantitative study of change possible. The Rasch model characterization of a person's performance on an item as a function of the difference between that person's ability and the difficulty of the item yields the useful result that sample-free item calibration, as well as test-free person measurement, can be obtained. There are many more reasons why a latent trait formulation is an important one (see, e.g., Bock & Wood, 1971; Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978; Lord & Novick, 1968; Rasch, 1960; Wainer, Morgan, & Gustafsson, 1979; Wright, 1968, 1977; Wright & Panchapakesan, 1969).

The problem in harvesting the benefits of latent trait models is the problem of fit, since these benefits follow only when the model fits. Studies of robustness (Lord & Novick, 1968, p. 492) indicate that certain parameters are robust with respect to modest deviations from the underlying assumptions; in particular, it seems that the Rasch model yields rather good estimates of ability and difficulty even when its assumption of equal slopes is only roughly approximated. The models that parameterize differential slopes have difficulty recovering the slope parameters even when the data do fit their model. Although this is not a topic of the present paper, it is desirable to indicate that attempts to expand the 1-parameter model to encompass additional possible characteristics of the data through an increase in the number of item parameters do not appear to be completely successful yet. Slope parameters are not well estimated in testing situations with only a few hundred individuals (Lord, 1979); and lower asymptotes, introduced to deal with guessing, cannot be consistently estimated (Ree & Jensen, 1979).

The Problem

If the Rasch model fits a given set of data, it has many practical benefits. It can never fit exactly, however, because there are always disturbances. These disturbances often take the form of (1) guessing, when a person of low ability gets a difficult item correct, and (2) sleeping, when a person of high ability gets an easy item wrong (Wright & Mead, 1977). The model has a certain amount of robustness with respect to such aberrations, but they can make the estimation procedures both biased and inefficient. The problem, then, is how to estimate the parameters of interest accurately and efficiently even when the data do not fit the model.

Some Choices

As a means of dealing with this problem, five different estimation schemes will be considered. These alternatives will be compared over a variety of simulations. It will be assumed that item difficulties are available and that only person abilities are to be estimated. This is a reasonable assumption because the calibration sample can be greatly increased. Individuals who have unusual patterns of response can be winnowed from it and a subset of individuals who are not "noisy" can be obtained. These individuals can then be used to obtain good estimates of item difficulty. However, the reverse is not true: A test of great length cannot be given, and when reporting on real persons, individuals who do not behave exactly as the model dictates cannot be eliminated. Abilities should be estimated for everyone. The task is to explore various estimation methodologies that assume the availability of item difficulties and to try to estimate ability as accurately and efficiently as possible. It may be that some of the techniques described will be of some use in the estimation of item difficulties as well, but this is not the primary motivation.

The Rasch Model

The Rasch model is based on the equation

$$P_{ij} = \exp(a_i - d_j)/[1 + \exp(a_i - d_j)] \qquad [1]$$

where

P_{ij} is the probability of person i answering item j correctly;

a_i is the ability of person i ($i = 1,\ldots,N$); and

d_j is the difficulty of item j ($j = 1,\ldots,L$).

Scheme 1: Pure Rasch

This is the standard maximum likelihood method for estimating Rasch abilities, given a vector of item difficulties. It relies on the Rasch model property that the raw score is a sufficient statistic for estimating ability. Each raw score has a distinct ability level associated with it. To find what it is, Equation 1 is solved for a_i, usually through Newton-Raphson:

$$r_i - \Sigma_j[p_{ij}] = 0, \qquad [2]$$

or

$$r_i - \Sigma_j[\exp(a_i - d_j)/(1 + \exp(a_i - d_j))] = 0 \qquad [3]$$

where r_i is the raw score for person i.

Scheme 2: Traditional Correction for Guessing

The traditional guessing correction is the assumption that if a person does not know the answer to a question and guesses, then the probability of guessing correctly is 1/M, where M is the number of choices. Thus, if there is an M-choice test and an individual has C wrong, it is assumed that he/she has an additional C/(M - 1) correct as a result of guessing. This is a crude attempt to put a lower asymptote on the item characteristic curve.

Scheme 3: Standard Jackknife

The Jackknife is an estimation scheme that was developed to reduce bias and has been shown (Tukey, 1958) to be useful for hypothesis testing as well. The way that it works in the application in this study is to construct a matrix of abilities, $\underset{\sim}{A}$, which has L - 1 raw scores labeling the rows and L + 1 columns. The first column, with elements $\underset{\sim}{A}(\underline{r}, 1)$ are the abilities associated with raw score $\underline{r}$, calculated through the method described in Scheme 1. The second column includes the ability levels based upon a test with the first item omitted. This test has only L - 1 items. Each succeeding column represents ability estimated through Scheme 1 but with that item omitted. Thus the $\underline{k}^{th}$ column is a test of length L - 1 containing all items except item $\underline{k}$ - 1.

The Jackknifed pseudovalues of ability are

$$a_j^* = LA(r,1) - (L-1)[x_j A(r-1, j+1) + (1-x_j)A(r, j+1)] \qquad [4]$$

where

$$\underline{x}_j = \begin{cases} 0 \text{ if item } \underline{j} \text{ is answered incorrectly} \\ 1 \text{ if item } \underline{j} \text{ is answered correctly; and} \end{cases}$$

the Jackknifed estimate of ability, $\underline{a}^*$, is just the mean of these $\underline{a}_j^*$'s:

$$a^* = \Sigma_j[a_j^*/L] = \mathrm{L}A(r,1) - [(L-1)/L] \; \Sigma_j[x_j \mathrm{A}(r-1, j+1) + (1-x_j)\mathrm{A}(r, j+1)] \qquad [5]$$

for $\underline{j}$ = 1, L.

For reasons that will become clear when the results of the simulations are discussed, it is important to note that the Jackknifed ability estimates are easy to compute. For any test all that has to be done is to compute the matrix $\underset{\sim}{A}$ and then, for each person, to run across the matrix at that person's raw score, adding up the entries in that row for each item that is incorrect and jumping up one row for each item that is correct. Jumping occurs when an item is correct, because the raw score for that person excluding that item is then one less.

Next, there are two aspects of an estimator that are of concern. First, it reduces bias, i.e., the effects of odd response patterns. The Jackknife was developed as a method to reduce bias (Quenouille, 1956), so it is hoped that it will serve this purpose. Secondly, it is desirable that the estimator does not fluctuate too much with minor disturbances in the response vector. This quality has been termed "resistance" (Tukey, 1977) and corresponds to an estimator having a sampling distribution with a small variance. The Jackknife is known to be modestly "resistant"; so this quality is likely to be met in practice as well.

To see how estimation with the Jackknife works, consider a test with 10 items whose difficulties are uniformly distributed, spanning a range of four logits. These difficulties are shown below:

-2.00	-1.56	-1.11	-0.67	-0.22
0.22	0.67	1.11	1.56	2.00

This yields the raw score-to-ability transformation matrix $\underset{\sim}{A}$, shown in Table 1.

Table 1
The Raw Score to Ability Conversion Matrix

Raw Score	Ability Estimate All Items	Ability Estimate Omitting Item *i*									
		1	2	3	4	5	6	7	8	9	10
1	-2.78	2.32	-2.45	-2.56	-2.63	-2.68	-2.72	-2.74	-2.75	-2.76	-2.77
2	-1.83	1.37	-1.45	-1.54	-1.63	-1.69	-1.74	-1.77	-1.79	-1.80	-1.81
3	-1.15	0.68	-0.73	-0.80	-0.88	-0.95	-1.01	-1.05	-1.09	-1.11	-1.12
4	-0.56	0.07	-0.10	-0.15	-0.21	-0.29	-0.35	-0.41	-0.46	-0.49	-0.51
5	0.00	0.51	0.49	0.46	0.41	0.35	0.29	0.21	0.15	0.10	0.07
6	0.56	1.12	1.11	1.09	1.05	1.01	0.95	0.88	0.80	0.73	0.68
7	1.15	1.81	1.80	1.79	1.77	1.74	1.69	1.63	1.54	1.45	1.37
8	1.83	2.77	2.76	2.75	2.74	2.72	2.68	2.63	2.56	2.45	2.32
9	2.78										

Consider how ability for a response vector of (1111110001) would be estimated. The raw score is 7, so the first 6 values associated with a raw score of 6 are summed (since the first 6 items were correct). Next, the three values (associated with Items 7, 8, and 9) associated with a raw score of 7 are added on, since these items were incorrect; so omitting them still yields a raw score of 7. Last, .68, the ability pseudovalue associated with a raw score of 6 for Item 10 omitted is added on. Summing these gives a total of 11.63. Next, this is multiplied by 9/10 [(L - 1)/L] and subtracted from 11.50 [L × 1.15], yielding a Jackknifed estimate for this person's ability of 1.03. Referring back to Table 1, it can be seen that a raw score of 6 yields an ability estimate of .56, which would have been the result if this person's answering the last item correctly had been treated as a wild guess and changed to incorrect. On the other hand, if this response were fully believed, his/her raw score would have been 7 and his/her ability estimate 1.15. The Jackknife weighs these two extremes and places the estimate between them.

Next, suppose that the response vector was (1111110010). Then, it is found that the pseudovalue of .68 associated with getting Item 10 correct is replaced with .73 (for Item 9) and 1.45 is replaced by 1.37. The net result of this changes the Jackknifed estimate from 1.03 to 1.06. This is just what a sensible person would do, since the second response pattern is more likely to have arisen through "proper" test taking and indicates a somewhat higher ability.

It appears that the Jackknife does what is desired, although how well is yet to be determined. It seems, however, from this demonstration that the variance of the sampling distribution of the Jackknifed ability is apt to be small, since large disturbances in response pattern do not cause large variations in the ability estimates. To see this, note that the ability estimate associated with the pattern (1111111001) is 1.09. (Other patterns can be attempted in order to observe how this estimation scheme behaves.) The Jackknife is not insensitive to response pattern, as Rasch estimates are, but it does not fluctuate much. This will be demonstrated in the results section.

Scheme 4: AMT-Robustified Jackknife

The pseudovalues obtained from standard Jackknifing suggest an additional estimation methodology. Consider the response pattern (1111110001) again. If the pseudovalue associated with each item is calculated using Equation 2, this gives

Item	Pseudovalue
1	1.42
2	1.51
3	1.69
4	2.05
5	2.41
6	2.95
7	-3.17
8	-2.36
9	-1.55
10	5.38

The mean of these pseudovalues yields the Jackknifed estimate of ability. Now consider these pseudovalues and how they are combined in the Jackknife. There are two kinds of pseudovalues--negative ones associated with incorrect responses and positive ones associated with correct responses. The Jackknife could be understood as first averaging the negative ones, thus coming out with an average ability estimate based upon items missed; then, averaging the positive ones for an ability estimate from the items answered correctly; and finally, combining these two averages, weighted by their sample sizes to yield the final Jackknifed estimate. It is known that the mean can be a poor way to estimate location. In some situations (Andrews, Bickel, Hampel, Huber, Rogers, & Tukey, 1972) it is the worst of all choices. Since concern is with unusual situations, perhaps the performance of the Jackknife can be improved through the choice of an estimator of location more robust than the mean.

Suppose the median of the positive pseudovalues is calculated. This is 2.05. The median of the negative pseudovalues is -2.36. Weighting these by 7 and 3, respectively, and summing and dividing by 10 yields an estimated ability of .73. Whether or not this is better than the Jackknifed value of 1.03 is difficult to determine, but it is certainly not too deviant.

One of the winners of the Princeton Robustness Study (Andrews et al., 1972) was the sine M estimator (the AMT). This estimator has an influence function nearly like that of the mean for observations in close but going to zero at the extremes. This implies that it will be efficient for nearly Gaussian distributions and robust against fat tails and outliers.

To understand how the AMT is calculated, consider that in regular cases, likelihood estimation of the location and scales parameters θ and σ of a sample from a population with known shape leads to equations of the form

$$\Sigma_j \; [-f'(z_j) \;/\; f(z_j)] = 0 \; , \qquad [6]$$

and

$$\Sigma_j \; [z_j f'(z_j) \;/\; f(z_j) - 1] = 0 \; , \qquad [7]$$

where f is the density function and $z_j = (x_j - \theta)/\sigma$.

M estimates of location are solutions, T, of an equation of the form

$$\Sigma_j \; \Psi[(x_j - T)/s] = 0 \qquad [8]$$

where Ψ is an odd function and s is estimated either independently or simultaneously.

The sine M estimate (AMT) is an M estimate in which the function ψ is

$$\Psi(x) = \begin{cases} \mathrm{Sin}(x/2.1) & |x| < 2.1\Pi \\ \\ 0 \text{ otherwise} . & \end{cases} \qquad [9]$$

The fourth scheme, then, is to use the AMT estimator on the positive and negative pseudovalues separately, obtaining two estimates of ability. These two estimates are then weighted by the number of observations that went into them and summed. The resulting value is then divided by the total number of items and the result is the AMT Jackknife estimate.

It is expected that when the test response pattern is reasonable (i.e., no responses are obtained that are unlikely, based upon the Rasch model), the AMT-Jackknife will look like the standard Jackknife. But when there are some odd responses, they will not be counted as heavily and thus will produce an estimate that is less affected by guessing and sleeping, while retaining the standard Jackknife's narrow sampling distribution.

Scheme 5: WIM

Wright and Mead (1976) developed a method for estimating ability in the Rasch model based upon an analysis of the residuals. Their method obtains an initial estimate of ability from raw score and its associated standard error, then calculates the residual of each item's response for that person by subtracting from the response the probability of its being correct. These residuals are standardized and a t statistic is calculated for the fit of this person's response pattern. If this t is greater than some chosen value (say, $t = 2$), then all items more than two logits above the person's initial ability estimate are omitted from that person's test and a new ability estimate is obtained based upon the shortened test. This process is repeated until an acceptable t is achieved or until the test becomes too short to work with.

This estimation scheme (WIM) was also included in the tests reported in this paper.[1] The results with this method reflect only on the method as it was received; there was no attempt to tune it by varying the critical t value. It could be that its performance would improve with fine tuning.

Method

The Guessing Model

How the individual responses in a simulation are characterized is critically important to its outcome. Certainly, if an estimator that matched the response generator was built, that estimator should emerge as superior in any competition. The validity of such investigations depends upon how the response model matches reality. It was decided that a reasonable model for responding has the following characteristics:

1. Need. A person guesses if he/she has a need to guess. This is a function of the extent to which the item is more difficult than the person is able. If people think they know the answer, they will not guess; if they do not, they might.

2. Invitation. This is a function of the item, unrelated to its difficulty (usually a function of the distractors). Some items invite guessing; others discourage it.

3. Inclination. This is a function of people unrelated to ability. Some people like to guess (risk takers?) and others do not (risk avoiders?).

4. Glitch. This represents something unexpected, which may be an item-person interaction unrelated to ability, difficulty, inclination, or invitation.

---

[1] The subroutine that performs WIM estimation was written by Ronald Mead.

The guessing model is

$$\Pi_{ij} = P_{ij} + (1 - P_{ij})(V_j + C_i - V_jC_i)/u_j \qquad [10]$$

where

π_{ij} is the probability of person *i* getting item *j* correct;

P_{ij} is the probability of person *i* getting item *j* correct based upon the Rasch model given earlier (the need to guess arises when P_{ij} is small because d_j is larger than a_i);

V_j is the invitation to guess associated with item *j* ($0 \leq V_j \leq 1$);

C_i is the inclination to guess associated with person *i* ($0 \leq C_i \leq 1$); and

u_j is the number of alternatives for item *j*.

The actual response that was generated by this model was determined. It was allowed to remain with probability 1 - G (where G is the glitch factor) and was changed with probability G, the generating parameter included to stir up trouble and add noise.

The Simulation

Independent variables. There are a large number of factors to be varied in a simulation in order to obtain a complete picture of what is happening. This simulation had eight factors that were systematically varied and on which all five estimation schemes were tried out. These were:

1. *Difficulty distribution* (3 levels). There were three distributions of difficulties that were used: uniform, Gaussian, and bimodal. The bimodal distribution was generated by constructing a uniform distribution and leaving out the middle half.

2. *Test length* (3 levels). Tests of three lengths were simulated: short (10 items), medium (20 items), and long (40 items). Longer tests were not used because the generalizability of results would increase only slightly but computer costs would multiply.

3. *Test width* (2 levels). Two test widths were simulated--narrow (2 logits) and medium (4 logits).

4. *Number of alternatives* (2 levels). Tests with five choices were simulated, since that reflects a common test format, as were tests with two alternatives (true-false format), which represents an extreme case.

5. *Ability* (4 levels). Four levels of ability were used: Very Low, Low, Medium, and High. Typically, Very Low was chosen as an ability that was the same as the easiest item on the test. Medium was typically chosen as zero, with Low halfway between them. High was usually symmetric with Low. Therefore, with the difficulties shown previously, the four abilities chosen would be -2, -1, 0, and +1. There was some variation in this choice, which will be explained below.

6. Invitation to guess (3 levels). This ranged from Low (6.0), to Medium (.5), to High (.9).

7. Inclination to guess (3 levels). The same as Invitation. As is evident from the response model, these two parameters are symmetric in their effect; so only the six interesting combinations were used.

8. Glitch (3 levels). Glitch is meant to convey rare, or at most seldom, trouble. Thus, three levels of glitch were used: none (.0), a little (.1), and a lot (.4). Note that a glitch of .5 is maximum, in that it will make the expected score for any response pattern the same (L/2).

Dependent variables. Two aspects of estimator performance were of interest. The first is accuracy: How different is the estimate of ability obtained from each estimator from the ability parameter that generated the response vector? This has been summarized by the mean difference between estimated ability for each estimator and the generating parameter. In the course of the simulation this was sometimes violated, because as a response vector was generated, it was checked to see if it was estimable. In particular, if a response vector had a raw score of 1 or lower or L - 1 or higher, it was not used, and another was generated. This resulted in a truncation of the ability distribution. This truncation caused the low-ability groups to have somewhat higher ability than the generating parameter would indicate and the high-ability group to have slightly lower ability than the generating parameter. To correct for this, the Rasch ability was estimated without any noise for a particular simulation situation (a specific length, width, distribution, and glitch) and the pure Rasch ability estimates were used as the basis of comparison for that simulation. Hence, when there was no noise, the Rasch estimates had zero bias by construction.

The second aspect of estimator performance that was of interest was the variance of the sampling distribution of that estimator around its own mean. Of course, the smaller this was, the better the estimator.

These two measures of estimator performance were combined into a total variance figure by adding together the weighted squared bias (analogous to the between-sum-of-squares) to the sampling variance (the within-sum-of-squares), using the usual synthesis of variance weightings. This represented the overall efficiency of each estimator. That estimator having the smallest efficiency for that sample was then found and each estimator's efficiency was divided into it to obtain relative efficiency. It is this figure that will be reported.

Results and Discussion

Obviously, with a design consisting of almost 4,000 cells and 5 estimators per cell, it would be impractical to attempt to present all the results. Instead, selected findings representative of the main effects will be presented, and some important interactions and trends will be discussed. The principle result was that one method was superior—the AMT-Jackknife. The AMT-Jackknife was superior, not because it was the most bias-free (although it did reasonably well in that regard), but rather because of its extremely small sampling variance.

No Noise

Before discussing the noisy simulations, the uncontaminated situation will be considered. It would seem that any estimation scheme proposed must do reasonably well in this situation before it an be considered a viable alternative to ordinary methods.

Table 2
Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = 0, Guessing Inclination = 0, Glitch = 0, for Five-Choice Items with a Uniform Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estimator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.7	.7	.7	.7	.8	.8	.9	.8	.9	.9	.9	.9
Traditional	.2	.1	.2	.2	.1	.1	.2	.4	.0	.1	.2	.3
Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
AMT-Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WIM	.7	.7	.7	.7	.8	.9	.9	.8	.9	.9	.9	.9
4 Logits												
Rasch	.8	.7	.7	.8	.8	.8	.8	.8	.9	.9	.9	.9
Traditional	.2	.2	.2	.3	.1	.1	.2	.4	.0	.0	.2	.3
Jackknife	1.0	.9	.9	1.0	1.0	.9	.9	.9	1.0	1.0	.9	.9
AMT-Jackknife	.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WIM	.7	.7	.6	.7	.7	.7	.8	.7	.7	.9	.9	.9

Table 2 shows the relative efficiencies (to 1 decimal place) of the 5 estimators for 3 test lengths, 2 different widths, and 4 abilities. The results for a uniform distribution of difficulties were striking for two reasons. First, they demonstrate the superiority of the AMT-Jackknife (followed closely by the standard Jackknife), assuring that the Jackknife is a viable scheme. Secondly, the Rasch maximum likelihood estimator was not the most efficient. This counters expectation, since maximum likelihood is supposed to yield estimates with minimum variance. Why did that fail to happen in this case? The answer is that the properties of maximum likelihood estimators are asymptotic. As test length increased, the relative efficiency of the Rasch estimator increased from 70% to 90%. The WIM estimator behaved in the same way. It would seem that 40 items is not enough for asymptotic properties to perform better than Jackknife properties. This finding leads to the reconsideration of the use of maximum likelihood estimators with short tests without further thought. Replacing maximum likelihood with AMT-Jackknife may benefit short test applications. The authors are not the first to observe that maximum likelihood does not accomplish everything desired from efficient estimation. Lewis (1970), in studying methods for the estimation of thresholds of sensitivity curves (a problem similar to the one being examined), found that maximum likelihood was unsatisfactory and used instead a scheme based on order statistics.

Table 3
Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = 0, Guessing Inclination = 0, Glitch = 0, for Five-Choice Items with a Gaussian Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estimator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.7	.7	.7	.7	.8	.9	.8	.8	.9	.9	.9	.9
Traditional	.2	.1	.2	.2	.1	.1	.2	.3	.0	.1	.2	.3
Jackknife	1.0	1.0	.9	.9	1.0	1.0	.9	1.0	1.0	1.0	1.0	1.0
AMT-Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WIM	.7	.7	.7	.7	.8	.8	.8	.8	.9	.9	.9	.9
4 Logits												
Rasch	.8	.7	.7	.7	.8	.8	.8	.8	.8	.9	.8	.8
Traditional	.1	.2	.2	.3	.1	.1	.2	.4	.0	.0	.2	.3
Jackknife	1.0	1.0	.8	.9	1.0	.9	.9	.9	1.0	1.0	.9	.9
AMT-Jackknife	.8	1.0	1.0	1.0	.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WIM	.7	.7	.6	.6	.8	.6	.8	.7	.9	.8	.9	.8

Table 3 shows that the same structure observed for a uniform distribution held for a Gaussian distribution. Once again, the AMT-Jackknife was superior, followed closely by the standard Jackknife, and then by Rasch and WIM. In all situations the Traditional guessing correction performed poorly. This was not unanticipated, since corrections are being made for a disturbance that is totally absent. As will be seen later, the performance of the Traditional estimator improved when guessing did occur (not surprisingly). Incidentally, WIM, which is the most computationally expensive procedure, is especially expensive for Gaussian and bimodal distributions of difficulty. More iterations are required for convergence in these situations than when the difficulties are uniform.

Table 4 shows the efficiencies for a bimodal distribution evidencing essentially the same structure that appeared with the other two distributions. WIM estimates were not obtained for a 40-item test (Width 2) when the procedure had not converged after 100 seconds (on an Amdahl/V6). It was felt that any information obtained from such a result would not be worth the cost or effort.

One conclusion is clear: When there is no guessing, the maximum likelihood estimator of ability in the Rasch model can be improved for tests of modest length (less than 40 items or so). In this noiseless situation there is little to choose from between the robust AMT-Jackknife and the standard Jackknife. The AMT was somewhat better but used a little more effort in its computation. It was also found that the Traditional correction for guessing, if applied when

Table 4
Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = 0, Guessing Inclination = 0, Glitch = 0, for Five-Choice Items with a Bimodal Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estimator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.6	.6	.5	.6	.8	.7	.6	.6	.9	.9	.6	.9
Traditional	.1	.1	.1	.2	.1	.1	.1	.2	.0	.1	.1	.2
Jackknife	.8	.7	.6	.8	1.0	.7	.6	.7	1.0	1.0	.6	1.0
AMT-Jackknife	1.0	1.0	1.0	1.0	.8	1.0	1.0	1.0	.8	1.0	1.0	1.0
WIM	.6	.6	.5	.6	.8	.6	.6	.6	.9	.9	.6	.9
4 Logits												
Rasch	.8	.6	.2	.8	.8	.9	.2	.9	.9	1.0	.2	.9
Traditional	.2	.1	.0	.2	.1	.1	.0	.3	.0	.1	.0	.2
Jackknife	1.0	.7	.2	.9	1.0	1.0	.2	1.0	1.0	1.0	.2	1.0
AMT-Jackknife	.6	1.0	1.0	1.0	.3	.7	1.0	.9	.2	.4	1.0	.5
WIM	.7	.6	.2	.7	.8	.8	.2	.8	*	*	*	*

guessing is absent, can have disastrous effects upon the efficiency of estimation. WIM worked as well as straight Rasch estimation when there was no guessing, although it did lead to some shrinkage due to the shortening of tests when unusual residuals occurred by chance.

Some Guessing

The next step in the exploration of estimators of ability was to study their behavior with a small amount of noise. Tables 5, 6, and 7 show the relative efficiencies for the three distributions with guessing invitations and guessing inclinations set at .5. Even a cursory examination shows that the structure observed in the no noise situation still obtained. The AMT-Jackknife and the standard Jackknife were still superior, but the WIM and the Traditional corrections improved. The bimodal distribution seemed to trouble the Jackknife more than its robustified version; however, both seemed to do satisfactorily. As would be suspected, at lower ability levels, schemes designed to deal with guessing (WIM and Traditional) worked to their best advantage. At higher ability levels, this was not the case. Jackknifing schemes did better on narrow tests than on wide ones, an observation that has been confirmed by examining their behavior on very wide tests of six to eight logits and noting a deterioration of performance; this was especially marked on eight logit-wide tests for the AMT.

The conclusions reached for noiseless data still hold, but less strongly. The two Jackknife methods remain the methods of choice, especially for individuals above mean ability. But as the data become increasingly noisy, each estimator reacted in its own way. The Rasch estimator yielded the same score for

Table 5

Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = .5, Guessing Inclination = .5, Glitch = 0, for Five-Choice Items with a Uniform Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estimator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.8	.8	.7	.6	1.0	.9	.9	.8	1.0	1.0	.9	.8
Traditional	.2	.3	.3	.4	.4	.5	.5	.5	.5	1.0	.7	1.0
Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.9
AMT-Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WIM	.8	.8	.7	.6	1.0	.9	.9	.8	1.0	1.0	.9	.8
4 Logits												
Rasch	.9	.8	.7	.6	.9	1.0	.8	.8	.8	1.0	.9	.7
Traditional	.4	.5	.3	.4	.6	.6	.7	.5	.4	1.0	.7	.8
Jackknife	1.0	.9	.8	.9	.9	1.0	.9	.8	.8	1.0	.9	.8
AMT-Jackknife	1.0	1.0	1.0	1.0	.9	1.0	1.0	1.0	.8	.9	1.0	1.0
WIM	.8	.8	.6	.5	1.0	1.0	.8	.6	1.0	1.0	.9	.7

Table 6

Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = .5, Guessing Inclination = .5, Glitch = 0, for Five-Choice Items with a Gaussian Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estimator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.8	.8	.7	.6	1.0	.9	.8	.7	1.0	.9	.8	.8
Traditional	.3	.3	.4	.4	.4	.4	.6	.5	.5	1.0	.7	1.0
Jackknife	1.0	1.0	1.0	.9	1.0	1.0	.9	.9	1.0	.9	.9	.8
AMT-Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
WIM	.8	.8	.7	.6	1.0	.9	.8	.7	1.0	.9	.9	.8
4 Logits												
Rasch	1.0	.9	.7	.5	1.0	1.0	.7	.6	1.0	.8	.8	.8
Traditional	.4	.5	.4	.3	.6	.7	.6	.4	.5	1.0	.7	.8
Jackknife	1.0	1.0	.9	.8	.9	1.0	.8	.8	1.0	.8	.8	.8
AMT-Jackknife	.9	1.0	1.0	1.0	.9	1.0	1.0	1.0	.9	.8	1.0	1.0
WIM	.8	.8	.7	.5	1.0	1.0	.7	.6	1.0	.9	.9	.8

Table 7
Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = .5, Guessing Inclination = .5, Glitch = 0, for Five-Choice Items with a Bimodal Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estmator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.8	.6	.5	.5	1.0	.8	.6	.5	1.0	.8	.6	.4
Traditional	.3	.3	.3	.2	.6	.6	.4	.3	.4	1.0	.4	.6
Jackknife	1.0	.7	.7	.7	1.0	.8	.6	.5	1.0	.8	.6	.5
AMT-Jackknife	1.0	1.0	1.0	1.0	.9	1.0	1.0	1.0	.8	.8	1.0	1.0
WIM	.8	.6	.5	.5	.9	.8	.6	.5	1.0	.8	.6	.4
4 Logits												
Rasch	.9	.6	.2	.7	.7	1.0	.2	.8	1.0	.8	.2	.5
Traditional	.4	.4	.1	.4	.4	.8	.2	.5	.6	1.0	.1	.6
Jackknife	1.0	.6	.2	.9	.7	1.0	.2	1.0	1.0	.8	.2	.5
AMT-Jackknife	.6	1.0	1.0	1.0	.4	.9	1.0	1.0	.4	.6	1.0	1.0
WIM	.8	.4	.2	.5	1.0	1.0	.2	.7	*	*	*	*

all raw scores of the same value, regardless of how that raw score was obtained, but yielded a poor goodness-of-fit statistic for misfitting persons. WIM reacted by shortening the test, indicating in essence that only a small portion of the test response vector obeys the Rasch model. The Jackknife methods reacted by regressing the scores toward zero (increasing bias but reducing variance of the sampling distribution) while increasing the standard error, thus signifiying that the information on the individual was small.

More Guessing

Next, the same three distibutions of item difficulty were considered, but this time with a great deal of guessing. Tables 8, 9, and 10 show the results when guessing invitation and inclination were both set to .9. This yielded a situation in which a person guessed whenever he/she did not know the answer and was identical to the situation posited in the derivation of the Traditional guessing correction. In this situation it would be expected that the Traditional method would excell; and it did perform well, but only when the test length was great enough to overcome its small sample inefficiency.

Once again, the same pattern of results emerged. For short tests the Jackknifing schemes worked best, with the edge always in the direction of the AMT. As tests got longer (40 items), the Traditional guessing correction began to work quite well. WIM, on the other hand, was disappointing, doing scarcely better than just a straight Rasch estimate. This must be interpreted, however. WIM reduces measurement bias quite well; but in doing so, it also decreases test length substantially. It could be argued that the length of the test evaluated

Table 8
Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = .9, Guessing Inclination = .9, Glitch = 0, for Five-Choice Items with a Uniform Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estimator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.8	.8	.7	.6	1.0	.9	.8	.8	.7	.5	.8	.6
Traditional	.4	.4	.4	.5	.7	.9	1.0	.8	1.0	1.0	1.0	1.0
Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	.9	1.0	.7	.5	.8	.7
AMT-Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	.9	1.0	.7	.5	.8	.7
WIM	.8	.8	.7	.6	1.0	.9	.8	.8	.7	.5	.8	.6
4 Logits												
Rasch	1.0	.9	.7	.6	.9	1.0	.6	.6	.6	.5	.7	.7
Traditional	.6	.8	.5	.4	.9	.9	1.0	.6	1.0	1.0	1.0	1.0
Jackknife	1.0	1.0	.8	.9	.8	1.0	.7	.7	.6	.5	.7	.8
AMT-Jackknife	1.0	1.0	1.0	1.0	.8	1.0	.8	1.0	.5	.5	.8	1.0
WIM	.8	.8	.6	.5	1.0	.9	.6	.5	.7	.6	.7	.6

Table 9
Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = .9, Guessing Inclination = .9, Glitch = 0, for Five-Choice Items with a Gaussian Distribution of Item Difficulties

	Test Length											
	10 Items				20 Items				40 Items			
Width and Estimator	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.9	.8	.7	.5	1.0	.9	.8	.7	.7	.5	.7	.6
Traditional	.5	.4	.4	.4	.6	.8	.9	.7	1.0	1.0	1.0	1.0
Jackknife	1.0	1.0	.9	.9	1.0	1.0	.9	.9	.7	.5	.8	.7
AMT-Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.7	.6	.8	.8
WIM	.9	.8	.7	.5	1.0	.9	.8	.7	.7	.5	.7	.6
4 Logits												
Rasch	1.0	.9	.6	.4	1.0	1.0	.6	.5	.6	.5	.6	.6
Traditional	.6	.9	.5	.4	1.0	.8	1.0	.5	1.0	1.0	1.0	.8
Jackknife	1.0	1.0	.8	.8	.9	1.0	.7	.7	.6	.5	.7	.6
AMT-Jackknife	1.0	1.0	1.0	1.0	.9	1.0	.8	1.0	.6	.5	.8	1.0
WIM	1.0	.8	.5	.4	1.0	.9	.6	.4	1.0	.9	.7	.5

by WIM, after eliminating items with large residuals, corresponds to the test that the testeee actually took. However, the reduced test length has the concomitant effect of increasing the standard error of measurement, and this causes its disappointing showing in the efficiency statistic.

Table 10
Relative Efficiencies of Five Estimators on Tests of Various Lengths and Widths for Four Ability Levels (Very Low, Low, Medium, High), with Guessing Invitation = .9, Guessing Inclination = .9, Glitch = 0, for Five-Choice Items with a Bimodal Distribution of Item Difficulties

Width and Estimator	Test Length											
	10 Items				20 Items				40 Items			
	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
2 Logits												
Rasch	.7	.6	.5	.4	1.0	.9	.6	.5	.6	.5	.6	.5
Traditional	.4	.4	.2	.4	.8	.8	.8	.4	1.0	1.0	1.0	.9
Jackknife	.9	.7	.6	.6	1.0	.9	.7	.6	.6	.5	.7	.6
AMT-Jackknife	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.6	.6	1.0	1.0
WIM	.7	.6	.5	.4	.9	.9	.6	.5	.6	.5	.6	.5
4 Logits												
Rasch	1.0	.6	.2	.4	.6	.9	.2	.5	.5	.4	.2	.3
Traditional	.9	.5	.2	.4	.7	1.0	.4	.6	1.0	1.0	.4	.5
Jackknife	1.0	.6	.3	.6	.6	.9	.2	.6	.5	.4	.2	.3
AMT-Jackknife	.9	1.0	1.0	1.0	.4	1.0	1.0	1.0	.2	.4	1.0	1.0
WIM	1.0	.5	.2	.4	1.0	.9	.2	.5	*	*	*	*

Guessing plus Glitching

Since the distribution of difficulties did not appear to have much effect on the behavior of the various estimators, the remainder of the results reported will be confined to one or the other of the distributions, with only side comments if the results differ substantially when another distribution was used. (Incidentally, for an extremely bimodal distribution in which all items are piled up at the extremes, the AMT will not work at all).

Table 11 shows the reaction of the various estimators to glitch of .1 over several test widths and for different amounts of guessing. There were no surprises. The deterioration of performance of the Jackknifing estimators with increased test width is visible but not severe. The AMT-Jackknife was always superior to the standard Jackknife. Under all conditions, Jackknifing seemed to be the best choice for higher ability individuals. Jackknifing also works rather well for correcting guessers, but the other methods may be better. In this table there are only reported results for test lengths of 20, but this is representative of the general findings. The Jackknifing methods did relatively less well with a test length of 40 and relatively better with a test length of 10.

Table 11
Relative Efficiencies of Various Estimators of Ability for a Test with 20 Items Whose Difficulties are Uniformly Distributed for Four Ability Levels (Very Low, Low, Medium, High), with a Random Noise Component of 10% (Glitch = .1)
(100 Entries Sampled Per Cell in Design)

Amount of Guessing (V,C) and Estimator	Test Width											
	2 Logits				4 Logits				6 Logits			
	Very Low	Low	Med.	High	Very Low	Low	Med.	High	Very Low	Low	Med.	High
(0,0)												
Rasch	1.0	.9	.9	.8	1.0	.8	.8	.9	.7	.9	.5	.9
Traditional	.2	.1	.2	.3	.4	.1	.2	.3	.8	.2	.1	.3
Jackknife	1.0	1.0	1.0	1.0	.9	.9	.9	1.0	.6	1.0	.6	1.0
AMT-Jackknife	1.0	1.0	1.0	1.0	.9	1.0	1.0	1.0	.4	.9	1.0	.9
WIM	1.0	.9	.9	.8	1.0	.6	.8	.8	1.0	.7	.3	.7
(.5,.5)												
Rasch	1.0	.9	.9	.8	.6	1.0	.8	.8	.4	.9	.6	.8
Traditional	.9	.5	.6	.4	1.0	.6	.6	.3	1.0	.7	.4	.4
Jackknife	1.0	1.0	1.0	1.0	.6	1.0	.9	.9	.4	.9	.6	1.0
AMT-Jackknife	1.0	1.0	1.0	1.0	.6	1.0	1.0	1.0	.3	.8	1.0	1.0
WIM	1.0	.9	.9	.8	.7	1.0	.8	.6	.8	1.0	.5	.6
(.9,.9)												
Rasch	1.0	.9	.9	.8	.7	.9	.8	.7	.3	.7	.6	.9
Traditional	.8	1.0	.7	.5	1.0	1.0	.7	.4	1.0	1.0	.5	.3
Jackknife	1.0	1.0	1.0	1.0	.7	.9	.9	.9	.3	.7	.6	1.0
AMT-Jackknife	1.0	1.0	1.0	1.0	.6	1.0	1.0	1.0	.3	.7	1.0	.8
WIM	1.0	.9	.9	.7	.8	1.0	.8	.8	.7	.9	.4	.6

True/False Tests

When the number of alternatives was reduced from five to two, much the same results were found. With no guessing the Jackknifing methods did best, with an edge to the AMT. As guessing became increasingly prevalent, the Traditional correction scheme worked better. It was still found, however, that for high abilities the AMT method was superior in efficiency to all others.

Standard Errors

The Rasch model standard error is

$$\text{Rasch (SE)} = 1/\{\Sigma_j[P_{ij}(1 - P_{ij})]\}^{\frac{1}{2}} \qquad [11]$$

for each ability level i. This accurately reflects what was observed empirically for the Rasch ability estimates in the simulations. When there was no guessing, the standard deviations of the sampling distributions was about what this equation would predict. It underpredicted the variability observed when there was noise. The WIM standard error is calculated in the same way as the Rasch except for a test of reduced length. This seems to accurately reflect reality for the situations tested.

The Jackknife standard error is calculated directly from the pseudovalues by

$$\text{Jackknife (SE)} = \left[\Sigma_j < a_j^* - a^* >^2 / < (L-1)L > \right]^{\frac{1}{2}} \qquad [12]$$

and is known to be a conservative estimator. This is certainly true in this case. It tended to overestimate the actual standard error by about 50% for test lengths of 10, by 25% for test lengths of 20, but was just about right for test lengths of 40.

Although there are several candidates for estimating the standard error of the AMT, the investigations of the authors are insufficient to be able to recommend one at this time. It seems reasonable to use the corrected Jackknife standard error until a better choice is found. The Jackknife standard error will almost certainly be conservatively large.

Conclusions

This investigation sought to find and test alternative methods for estimating ability under the Rasch model in the face of plausible noise. This was done by using some recent developments in robust estimation without adding parameters to the model, thus retaining the Rasch model's attractive attributes. It was found that gains in recovering abilities in the presence of guessing and untoward responses of other kinds can be obtained through the use of a robustified Jackknife. But it was also found that specially developed models aimed at the lower end of the ability continuum may be able to accomplish this better than these general tools. WIM worked when there was guessing and aided in increasing the accuracy of estimation for low-ability testees. The Traditional method worked when there was much guessing, the test was long, and the ability of the testees was low.

A surprising finding was that for short tests of 10 or 20 items, the Jackknife estimators, with a significant edge to the AMT version, yielded better estimates of ability than the maximum likelihood estimator, even when preconditions for the Rasch model held. This increase in efficiency of estimation is especially important for those applications of latent trait models that use a limited number of measures obtained about a person as a de facto test (see, e.g., the analysis of parole data in Perline, Wright, & Wainer, 1979). In these circumstances the number of items cannot be easily increased, and the only alternative is to improve the estimate of ability through other means. Thissen (1976) attempted to do this by using a method Bock (1972) developed for wrong answers, but this is very expensive computationally and only applicable to multiple-choice items. Super-efficient estimators may also be useful in such applications as adaptive testing.

The simulations performed were very extensive; nevertheless, considerably more research is necessary. A careful study of estimators of standard error is critical, as are the distributional properties of the Jackknifed estimators. Robust estimators have not been used in conjunction with the Jackknife before, so nothing is known about that distribution. The authors believe that Jackknife estimates are t-distributed (although there is difficulty in determining the effective degrees of freedom). It seems reasonable, therefore, to suppose that the robust Jackknife will have a similar symmetric (albeit tighter) distribution. This suggests that the Jackknife estimates of standard error for the AMT estimator are conservative. Just how conservative these actually are, however, awaits further investigation.

A second area of investigation that is still incomplete is goodness-of-fit tests. Substituting robust estimates of ability into the usual goodness-of-fit equations should yield a conservative estimate more realistic than those usually obtained (which benefit from capitalization on chance). But it is not known to what extent the asymptotic properties of such fit statistics derived and/or described by Anderson (1973), Fischer (1974), Martin-Löf (1974), and Wright and Stone (1979) apply.

The finding of improved estimation efficiency is an intriguing one. Lewis (1970) pointed out that although maximum likelihood estimates of location parameters of ogive functions are asymptotically identical to minimum chi-square estimates, they can be quite different for small samples. Neither makes any claims for small sample efficacy, but what is surprising is how large "small" can be and how much of an improvement can be made using an alternative procedure. Lewis found that asymptotically optimal procedures did especially poorly in estimating accurate confidence intervals around the location parameter. Perhaps this, too, is an area in which the AMT-Jackknife will prove useful. The questions are clear and important, and the methodology for answering them is straightforward.

There are a number of other estimators that may improve performance still more. For example, Ramsay (1977) found that the E_a estimator has some advantages over the AMT. Novick (1979) has suggested several Bayesian estimators that may have promise.

The main finding of this study is that for short tests the asymptotic properties of maximum likelihood estimators are not fully realized. Other methods increase efficiency. In addition, these other estimators can correct for noise in the data, such as guessing, and thus can increase validity. The AMT-Jackknife may not be the best estimator of its type that can be derived. Perhaps other variations on this theme can go even further in the direction of super-efficiency. Nevertheless, the AMT-Jackknife does seem to deal well with the problem of guessing, which is so poorly handled by estimation of a lower asymptote of the item characteristic curve.

REFERENCES

Anderson, E. B. A goodness-of-fit test for the Rasch model. *Psychometrika*, 1973, *38*, 123-140.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. *Robust estimates of location*. Princeton, NJ: Princeton University Press, 1972.

Bock, R. D. Estimating item parameters and latent trait ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29-51.

Fischer, G. H. *Einfurung in die theorie psychologischer tests. Grundlagen und anwendungen*. Bern: Huber, 1974.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 1978, *48*, 467-510.

Lewis, C. The countback method (ETS RB 70-30). Princeton, NJ: Educational Testing Service, 1970.

Lord, F. M. Small N justifies Rasch methods. In D.J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1980.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Martin-Löf, P. Exact tests, confidence regions, and estimates. *Memoirs, No. 1*. Aarhus, Denmark: University of Aarhus, Institute of Mathematics, Department of Theoretical Statistics, 1974. (Proceedings of Conference on Fundamental Questions in Statistical Inference).

Novick, M. R. Discussion. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1980.

Perline, R., Wright, B. D., & Wainer H. The Rasch model as additive cojoint measurement. *Applied Psychological Measurement*, 1979, *3*, 237-255.

Quenouille, M. Notes on bias in estimation. *Biometrika*. 1956, *43*, 353-360.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danske Paedagogiske Institut, 1960.

Ramsay, J. O. A comparative study of seveal robust estimates of slope, intercept, and scale in linear regression. *Journal of the American Statistical Association*, 1977, *72*, 608-615.

Ree, M. J., & Jensen, H. E. Effects of sample size in the estimation of item parameters. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory, 1980.

Thissen, D. M. Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, 1976, *13*, 201-214.

Tukey, J. W. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 1958, *29*, 614. (Abstract)

Tukey, J. W. *Exploratory data analysis*. Reading, MA: Addison-Wesley, 1977.

Wainer, H., Morgan, A., & Gustafsson, J. E. *A review of estimation procedures for the Rasch model with an eye toward longish tests*. Manuscript submitted for publication, 1979.

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1968.

Wright, B. D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 97-116.

Wright, B. D., & Mead, R. J. *Analysis of residuals*. Unpublished manuscript, 1976.

Wright B. D., & Mead, R. J. *The use of measurement models in the definition and application of social science variables* (Technical Report DAHC19-76-G-0011). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 1977.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-48.

Wright, B. D., & Stone, M. H. *Best test design*. Chicago: MESA Press, 1979.

ACKNOWLEDGMENTS

This research was funded through a grant from the Law Enforcement Assistance Administration (78-NI-AX-0047) to the Bureau of Social Science Research, Howard Wainer, Principal Investigator. We thank Ronald Mead, Anne Morgan, and James Ramsay for their kind, generous, and invaluable help at various stages of the project.

APPROPRIATENESS MEASUREMENT: BASIC PRINCIPLES AND VALIDATING STUDIES

MICHAEL LEVINE
UNIVERSITY OF ILLINOIS

FRITZ DRASGOW
YALE UNIVERSITY

In a large test administration a few examinees may be so unlike other examinees that their multiple-choice aptitude test scores have limited value as ability measures. A particularly transparent example is provided by a hypothetical low-ability copier who copies half of his/her answers from a much more able neighbor. Other test anomalies include:

1. Improperly coached examinees who are shown the answers to some items before the exam begins,
2. Examinees with high ability but atypical schooling or low English fluency,
3. Exceptionally creative examinees who discover novel interpretations for some items,
4. Examinees who are very conservative in their use of partial information, and
5. Examinees who make alignment errors on their answer sheet over a block of items, answering the 9th item in the 10th place, the 10th item in the 11th place, and so forth.

In each of these cases it can be argued that (1) the test score is not an appropriate measure of ability and (2) the item-by-item pattern of answers may be recognizably unusual. For example, the hypothetical low-ability copier seems likely to have many easy items incorrect and many difficult items correct, relative to typical examinees. A second example of an inappropriate test score and an unusual answer pattern occurring together is provided by the hypothetical examinee with an alignment error. He or she will most likely have a block of consecutive items incorrect and an unusual answer pattern within the block.

Thus, a multiple-choice aptitude test may be a dubious measure of ability due to any one of many (although possibly rare) causes. In at least some cases, the item-by-item response pattern may contain evidence of this fact. This paper considers the problem of using answer patterns to recognize inappropriate test scores.

Appropriateness Measurement: Its Objectives and Limitations

Appropriateness measurement is a general approach to the problem caused by inappropriate test scores. Its purpose is simply to identify inappropriate test scores. It is limited to cases, such as those noted above, in which inappropriate test scores and unusual answer patterns tend to co-occur. Appropriateness measurement is implemented by statistics, called appropriateness indices, that measure the degree to which an examinee's answer pattern is "unusual," i.e., unlike the pattern expected from typical examinees.

In appropriateness measurement studies, examinees are sorted into two groups: (1) examinees with very unusual answer patterns, as indicated by very extreme index values and (2) other examinees, i.e., examinees with typical index values. Appropriateness measurement is successful to the extent that the group of examinees with extreme index values has a larger proportion of examinees with inappropriate scores than the group with typical index values.

Background

The first large-scale appropriateness measurement study was reported by Levine and Rubin (in press). This study, reviewed below, provides the background and context for the theoretical developments and empirical results reported in this paper.[1]

Levine and Rubin (in press) identified three types of appropriateness indices and reported positive empirical findings with these indices. However, the generality of their findings is limited by properties of their data set. In particular, their data were simulated, the simulation parameters were available for use in defining appropriateness indices, and aberrant examinees were unequivocally identified.

In this paper actual and simulated data are used to attack three problems raised by the Levine and Rubin study, namely:

1. Estimated versus known item parameters. With simulated data, item parameters (e.g., item difficulties) are known and need not be estimated prior to computing appropriateness indices. With actual data, parameters must be estimated. How seriously will appropriateness measurement be affected by estimation errors?

2. Unidentified aberrants. In a simulation study, atypical examinees are unequivocally identified and a sample of truly normal examinees is available for item parameter estimation. With actual data an unknown proportion of unidentified aberrants will be included in each large sample of nominally normal examinees. How will the presence of these aberrants affect parameter estimation and, consequently, appropriateness measurement?

---

[1] For independent contemporary work that appears similar in conception, see Flier (1977). For possibly related research based on classical test theory, see Ghiselli (1956; 1960a; 1960b) and Donlon and Fischer (1968).

3. *Model validity*. Simulated data conform precisely to the psychometric model used to generate data and to formulate appropriateness indices. There will be reliable contradictions, however, to the assumptions of any tractable, nontrivial psychometric model in a large sample of actual data. Are currently available psychometric models sufficiently valid to support appropriateness measurement with actual data?

The results presented will show (1) that appropriateness indices are not seriously degraded by the use of estimated parameters; (2) that appropriateness indices are not seriously degraded even when a relatively large proportion of aberrant examinees is initially (and improperly) treated as normal; and (3) that detection rates with actual test data are comparable to detection rates with simulated data.

Review of Test Theories and Basic Appropriateness Measurement

Appropriateness measurement involves a two-stage process: a test norming or item parameter estimation stage, followed by a person measurement or index computation stage. This distinction parallels the separation of parameters in latent trait models into (1) item or test characterizing parameters, such as item difficulties; and (2) individual difference or person characterizing parameters, such as abilities.

The Standard Model

In the studies to be reported here, the test norming stage is developed around what will be called the *standard model*. This test model is a version of the 3-parameter logistic model of item response theory (Birnbaum, 1968). According to the standard model, an answer sheet is generated by a two-stage experiment. In the first stage an ability, θ, is sampled. The second stage is a sequence of independent binary random variables. These are the item scores, coded from the observed answer sheet with "1" denoting a correct response and "0" an incorrect response. (The ability θ is not observed and the distribution of abilities is neither specified nor estimated in these studies.)

After some notation is introduced, the essential features of the standard model can be summarized with two equations. Let $\underset{\sim}{U}^{(j)}$ denote the $\underline{j}^{th}$ examinee's answer pattern. Thus, $\underset{\sim}{U}^{(j)}$, $\underline{j} = 1, 2, \ldots, N$, is the vector of item scores,

$$\underset{\sim}{U}^{(j)} = < u_1^{(j)}, u_2^{(j)}, \ldots, u_n^{(j)} > ,$$

for a test composed of $\underline{n}$ items. Let $P_i(\theta)$ denote the regression of the $\underline{i}^{th}$ item score on ability, i.e., $P_i(\theta) = (\underline{u_i}|\theta)$. For a given θ, $P_i(\theta)$ can be interpreted as the conditional probability of an examinee, randomly selected from all examinees with ability θ, correctly answering the $\underline{i}^{th}$ item. With this notation, the conceptualization of the second stage of the answer sheet generation process can be expressed by the following equation, which is known as the "local independence" assumption of latent trait theory:

$$\text{Prob}\,\{\underset{\sim}{U}^{(j)} \mid \theta^{(j)} = \theta\} = \prod_{i=1}^{n} P_i(\theta)^{u_i}\,[1 - P_i(\theta)]^{1-u_i} \qquad [1]$$

In words, Equation 1 states that item responses are conditionally independent. The independence of different examinees implicit in the first stage of the answer sheet generation process is expressed in Equation 2:

$$\text{Prob}\,\{\underset{\sim}{U}^{(1)}, \underset{\sim}{U}^{(2)}, \ldots, \underset{\sim}{U}^{(N)} \mid \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\} \qquad [2]$$

$$= \prod_{j=1}^{N} \text{Prob}\{\underset{\sim}{U}^{(j)} \mid \theta^{(j)}\}\ .$$

To facilitate the test norming stage, each P_i is assumed to have the functional form

$$P_i(\theta) = c_i + (1 - c_i)\,\{1 + \exp[-a_i(\theta - b_i)]\}^{-1} \qquad [3]$$

for some positive $\underline{a_i}$, real $\underline{b_i}$, and $\underline{c_i}$ with $0 \leq \underline{c_i} \leq 1$. This asserts that P_i is S-shaped with a lower asymptote of $\underline{c_i}$ and an upper asymptote of unity. The location and scale parameters $\underline{b_i}$ and $\underline{a_i}$ express differences between items in difficulty ($\underline{b_i}$) and ability to discriminate between low- and high-ability examinees ($\underline{a_i}$). This particular functional form is conventional, is widely used, and has been supported in nonparametric studies (Levine & Saxe, 1976).

"Test norming" consists of estimating the numerical values for $\underline{a_i}$, $\underline{b_i}$, and $\underline{c_i}$. This is done by selecting a large sample of N presumably normal examinees; observing their answer patterns $\underset{\sim}{U}^{(1)}$, $\underset{\sim}{U}^{(2)}, \ldots, \underset{\sim}{U}^{(N)}$; and finding a set of $\underline{a}$'s, $\underline{b}$'s, $\underline{c}$'s, and θ's that maximizes the likelihood function in Equation 2. The interest in the test norming stage is in the item parameters $\underline{a}$, $\underline{b}$, and $\underline{c}$; however, the θ's, must also be estimated in the current procedure.

Lord's LOGIST algorithm (Wood & Lord, 1976; Wood, Wingersky, & Lord, 1976) can be used to maximize Equation 2. This program has been vigorously criticized by Wright and his associates (e.g., Wright, 1977). In fact, Wright has questioned whether any algorithm can be designed to estimate the parameters of the standard model; the relevance of the results of the present studies to these criticisms is summarized in the last section of this paper.

The test norming stage can be thought of as specifying a continuum of models or statistical characterizations of typical examinees by using test responses of the first N nominally normal examinees in the first stage to estimate a set of item parameters. These parameter estimates can be substituted in Equation 1 to obtain an explicit formula for the likelihood of a new pattern of answers, say $\underset{\sim}{U}^{(N+1)}$. An intuition that had guided much of the authors' current research, and which will now be used to introduce the person measurement stage is this: Suppose for all values of θ, $\underset{\sim}{U}^{(N+1)}$ appears improbable in the sense that Prob($\underset{\sim}{U}^{(N+1)}|\theta$) is very small. Then, $\underset{\sim}{U}^{(N+1)}$ is badly fit by all models for individual data developed in the test norming stage.

In the person measurement stage the item parameters are treated as known and an index of goodness of fit is computed for each person's answer pattern. The simplest index, L_0, is

$$L_0 = \log \max_{\theta} \text{Prob}\ (\underset{\sim}{U}^{(N+1)}|\theta) \ . \qquad [4]$$

This index will be small if Prob($\underset{\sim}{U}^{(N+1)}|\theta$) is small for values of θ. A small value of L_0 could result if many incorrectly answered easy items rule out high values of θ and many correctly answered difficult items rule out low values of θ. L_0 detects aberrance surprisingly well. To improve upon it, a model for aberrant data and two measures for the degree of aberrance will be specified in the next section.

Variable Ability Models and Appropriateness Indices

L_0, like χ^2, is sensitive to any type of poorness of fit for the standard model. A generalization of the standard model is needed to detect aberrations of the specific kind that is of interest.

In many of the most important types of test inappropriateness, the aberrant examinee behaves as if his or her ability were fluctuating from item to item. Thus, the low-ability cheater appears to have a much higher ability for those items on which he or she has been coached. The high-ability, low-English-fluency candidate behaves as if he or she had low ability on linguistically demanding items.

In the standard model the examinee's ability, θ, is constant across items. In variable ability models, introduced by Levine and Rubin (1979), the examinee's ability is conceptualized as varying from item to item. For example, in the Gaussian model, each examinee is characterized by a pair of parameters: central ability, θ_0, and ability variance, σ^2. (Note that the standard model utilizes a single individual difference parameter, θ.) According to the Gaussian model, the examinee has ability θ_1 on Item 1, θ_2 on Item 2,..., and θ_n on Item n. The model asserts that for given θ_0 and σ^2, the θ_i are independent identically distributed normal or Gaussian random variables with mean = θ_0 and variance = σ^2. To specify the model more precisely the likelihood function Prob($\underset{\sim}{U}|\theta_0, \sigma^2$), which gives the conditional probability of a vector $\underset{\sim}{U}$ of item responses, is written and simplified as follows:

$$\text{Prob}\ (\underset{\sim}{U}|\theta_0, \sigma^2) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} \prod_{i=1}^{n} \{P_i(\theta_i)^{u_i} \qquad [5]$$

$$[1-P_i(\theta_i)]^{1-u_i}\ \phi(\theta_i;\theta_0,\sigma^2)d\theta_i\}$$

$$= \prod_{i=1}^{n} \int_{-\infty}^{\infty} P_i(t)^{u_i}[1 - P_i(t)]^{1-u_i} \phi(t;\theta_0,\sigma^2)\,dt \quad , \qquad [5]$$

where $\phi(\underline{t};\, \theta_0,\, \sigma^2)$ is the Gaussian density

$$\phi(t;\theta_0,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}\left[\frac{t-\theta_0}{\sigma}\right]^2\right\} . \qquad [6]$$

Equation 5 is analogous to Equation 1 for the standard model.

To obtain a second appropriateness index testing for a specific departure from the standard model, the maximum likelihood ratio test statistic is computed:

$$LR = L_n - L_0 \qquad [7]$$

where

$$L_n = \log \max_{\theta_0,\sigma} \text{Prob}(\underset{\sim}{U}|\theta_0,\sigma^2)$$

and

$$L_0 = \log \max_{\theta} \text{Prob}(\underset{\sim}{U}|\theta),$$

as before. LR measures the degree to which a variable ability model provides a better fit to the observed pattern of responses than the standard model.

The final appropriateness index or measure of goodness of fit is $\hat{\sigma}$, the maximum likelihood estimate of the ability standard deviation. This index is obtained by maximizing $\text{Prob}(\underset{\sim}{U}|\theta_0, \sigma^2)$ with respect to both θ_0 and σ. The standard model is a special case of the Gaussian model in the sense that $\text{Prob}(\underset{\sim}{U}|\theta)$ is the limit of $\text{Prob}(\underset{\sim}{U}|\theta_0, \sigma^2)$ as σ decreases to zero. Consequently, a small $\hat{\sigma}$ can be interpreted as indicating a small degree of aberrance.

Index Evaluation and Receiver Operator Curves

Each application of appropriateness measurement will have different rewards for correctly identifying an inappropriate score and different penalties for incorrectly classifying a nonaberrant examinee as aberrant. The receiver operator curve (ROC) of statistical decision theory provides a graphic way to compare indices prior to the specification of rewards and penalties.

In applying an appropriateness index to classify examinees, a cutoff or criterion value of the index is specified. For the present exposition, assume that small index values indicate aberrance. At each criterion value, $\underline{t}$, the proportion of aberrant examinees correctly identified as aberrant and the proportion of normal examinees improperly identified as aberrant can be denoted as

$\underline{x}(\underline{t})$ = proportion of normal examinees with index values $< \underline{t}$
$\underline{y}(\underline{t})$ = proportion of abberrant examinees with index values $< \underline{t}$.

An ROC results from plotting the $< \underline{x}(\underline{t}), \underline{y}(\underline{t}) >$ pairs obtained for various values of $\underline{t}$. A desirable ROC is one that rises sharply from the origin toward the upper left-hand corner of the plot. In contrast, an ROC that lies along the diagonal of the plot indicates a random classification rule. That is, classifying examinees or the basis of flipping a coin yields a diagonal ROC. Clearly, an ROC that indicates an effective detection procedure is one that lies well above the diagonal.

Simulation Procedures and Results

Levine and Rubin's (in press) simulation methods and results are reviewed here, as some of their data is reanalyzed, and their methods are applied to new data. To simulate a normal vector of item scores, they first sampled an ability θ from a normally distributed population with zero mean and unit variance. The examinee's first item score was generated by sampling a number uniformly distributed in the unit interval. If the sampled number was less than or equal to $P_1(\theta)$ from Equation 3, then the first item was scored as correct; otherwise, the item was scored as incorrect. The remaining item scores were obtained by independently drawing new uniformly distributed numbers and comparing them with $P_i(\theta)$ for $\underline{i} = 2, 3, \ldots, \underline{n}$.

The parameters $\underline{a}_i$, $\underline{b}_i$, and $\underline{c}_i$ utilized in Equation 3 were those obtained by Lord's fitting of a 3-parameter logistic model to a large sample of Scholastic Aptitude Test, Verbal Section data (SAT-V; Lord, 1968). The actual simulation was implemented with Hambleton and Rovenelli's (1973) program. (For technical details concerning the random number generators used, see Levine & Rubin, in press, Appendix).

Aberrant examinees were simulated by modifying simulated normal answer sheets in various ways. In this paper concern is primarily with the "20% spuriously low" modification, but other modifications will also be reviewed briefly.

To create a spuriously high answer sheet, 20% of a normal simulated examinee's item responses were randomly sampled without replacement. If the sampled item was originally scored as correct, it is left unchanged. If the sampled item was not correct, it is rescored as correct.

To create a spuriously low answer sheet, 20% of a normal answer sheet's responses were sampled as before. Then, a random number generator was used to simulate random guessing over the five multiple-choie alternatives. No matter what the original sampled item score was, the sampled item was scored correct with probability .20 and scored as incorrect with probability .80.

Levine and Rubin observed that the modification of a normal answer pattern to create a spuriously low answer pattern frequently resulted in little or no change in the actual response vector. Clearly, if an aberrance-producing process does not alter the objective response pattern, it is futile to attempt to detect the presence of the process with an appropriateness index. Furthermore, if there is no effect on the objective response pattern and test score, there is little motivation for detecting aberrance. Consequently, Levine and Rubin separately analyzed the data from spuriously low examinees who had at least 10% of their scores changed. Both correct responses changed to incorrect and incorrect responses changed to correct were counted toward the 10% figure.

Levine and Rubin observed good aberrance detection with the total sample of spuriously low examinees and excellent detection for both the spuriously high examinees and the selected large-score-change sample of spuriously low examinees. These generalizations held for all three indices.

There are two additional results obtained by Levine and Rubin that merit attention. First, Levine and Rubin systematically varied the percentage of items modified, utilizing 4%, 10%, 20%, and 40% treatments. As expected, they found increasing detectability as the percentage of modified items was increased. A second finding by Levine and Rubin was that the three appropriateness indices--L_0, LR, and $\hat{o}$--yield quite similar patterns of detectability. Interestingly, no one index was substantially better or worse than the other two. Consequently, Levine and Rubin did not recommend using any one particular index in future research.

For the present study Levine and Rubin's data were reanalyzed using estimated item parameters instead of simulation parameters, and appropriateness measurement techniques were applied to actual test data. Levine and Rubin data files relevant to the present research were as follows:

1. NORMAL 3200: 3,200 simulated answer sheets with normally distributed abilities; item parameters from Lord's (1968) fitting of the SAT-V; items scored either as correct or incorrect.

2. NORMAL 2800: The first 2,800 records from NORMAL 3200.

3. LOW 200: Records 3,001, 3,002,...,3,200 from NORMAL 3200 modified to simulated ability-unrelated responding on 20% of the test according to the 20% spuriously low modification.

4. LOW 102: 102 records selected from LOW 200, having at least 10% of the simulated examinee's original responses changed in the spuriously low modification.

5. HIGH 200: Records 2,801, 2,802,...,3,000 from NORMAL 3200 modified to simulate cheating on 20% of the items according to the 20% spuriously high modification.

Study 1: Estimated Parameters

Problem

Levine and Rubin bypassed the first (test norming) stage of appropriateness measurement. Instead of estimating item parameters from a large sample of normal examinees, they used the exact (simulation) parameters to compute appropriateness indices. Clearly, in an application with actual data, item parameters must be estimated. How sensitive are indices to item parameter estimation error? Can high detection rates be achieved with estimated parameters?

Method

Item parameters were estimated by applying Lord's maximum likelihood algorithm LOGIST to NORMAL 2800 data. L_0 was computed for each NORMAL 2800 examinee by evaluating the likelihood function at the LOGIST maximum likelihood estimate of ability. The LOGIST-estimated item parameters were then used to compute L_0 for the LOW 102 response vectors by rerunning LOGIST for the LOW 102 data with all item parameters fixed at the values obtained from NORMAL 2800. In this way an estimated parameter L_0 appropriateness index value was obtained for each NORMAL 2800 and LOW 102 response vector.

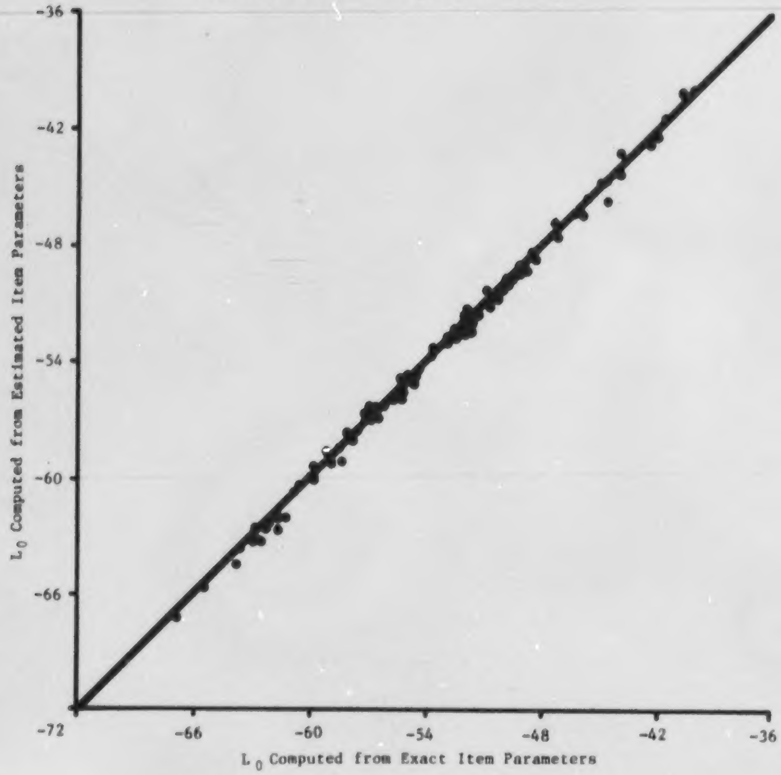
Results

A close agreement between values of L_0 computed from exact parameter and estimated parameter index values was observed. In Figure 1 a bivariate scatterplot demonstrates this agreement for the critical LOW 102 sample. Each of the 102 simulated examinees contributes a point to the scatterplot. The *x*-coordinate of the point is L_0 computed with exact item parameters; the *y*-coordinate, with estimated item parameters. If there were perfect agreement between the two measures, the points would fall on the diagonal line, which has been drawn on the figure. A very slight tendency is observed for estimated index values to be smaller than exact index values for the most aberrant examinees in LOW 102 (L_0 less than -60).

The same close agreement was observed for the NORMAL 2800 sample. This is shown in Figure 2 for the first 100 simulated examinees in NORMAL 2800.

A more significant result is shown in the estimated parameter L_0 ROC presented in Figure 3. Note that detection rates were high, even for low false alarm rates. For example, 12.7% of the aberrant examinees could be correctly classified with an L_0 cutoff score that did not misclassify a single NORMAL 2800 examinee. Further, 21.6% of the aberrant examinees were detected at a false alarm rate of .9% and 47.1% of the aberrants were detected at a 4.4% false alarm rate. Figure 3 is in close agreement with Levine and Rubin's exact parameter LR ROC computed from the same data (Levine & Rubin, in press, Figure 8). In fact, there is a very small superiority for the estimated parameter ROC; using estimated parameters there were 0%, .9%, and 4.4% false alarms at hit rates of 11.8%, 21.6%, and 47.1% in comparison to .1%, .9%, and 4.8% false alarms at the corresponding points in the exact parameter ROC.

Figure 1
Bivariate Scatterplot of L_0 Computed from Exact and Estimated (from NORMAL 2800 Data) Item Parameters for LOW 102 Response Vectors



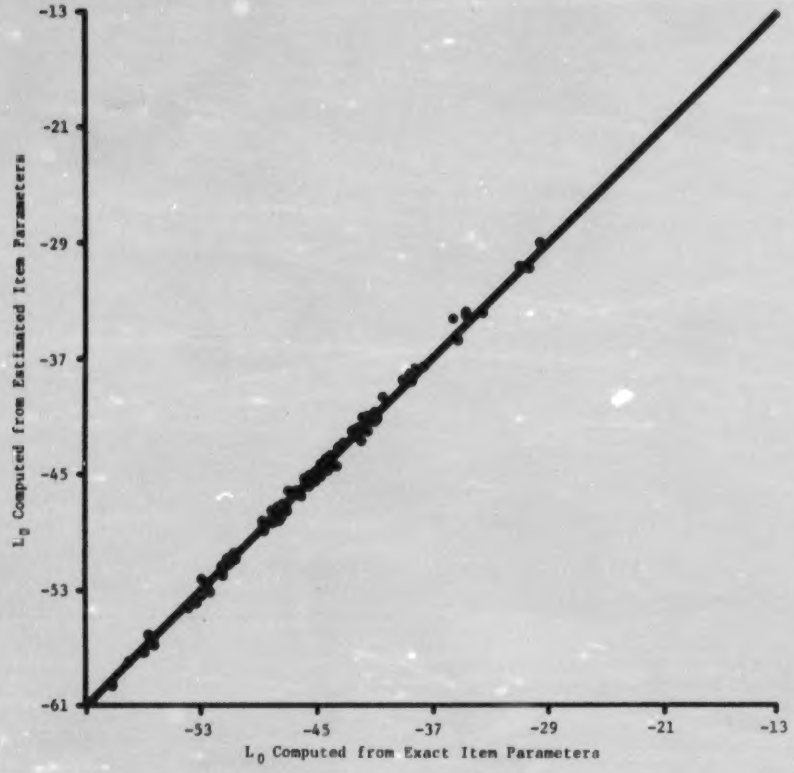
Discussion

The results may initially seem surprising in view of Lord's (1975a) study of the disparity between LOGIST-estimated item parameters and simulation parameters.[2] Lord found much more variation around the diagonal in his bivariate plots of simulation item parameters versus estimated item parameters than appears in the plots of estimated parameter L_0 versus simulation parameter L_0. The discrepancy becomes less surprising when it is recalled that L_0 is the maximum value of a likelihood function, whereas a LOGIST parameter estimate gives the location of a point at which the maximum is obtained. If the likelihood function is relatively flat, then the maximizing values of the arguments of the function will be difficult to determine precisely because a small parameter change results in a very small likelihood change. Somewhat paradoxically, flatness of the likelihood function can simplify the problem of calculating L_0, the value of the function near its maximum. The value of the likelihood function will be almost constant for parameter values in the neighborhood of the maximizing value.

---

[2] LOGIST version 2.B was used in the present research. LOGIST has since been modified.

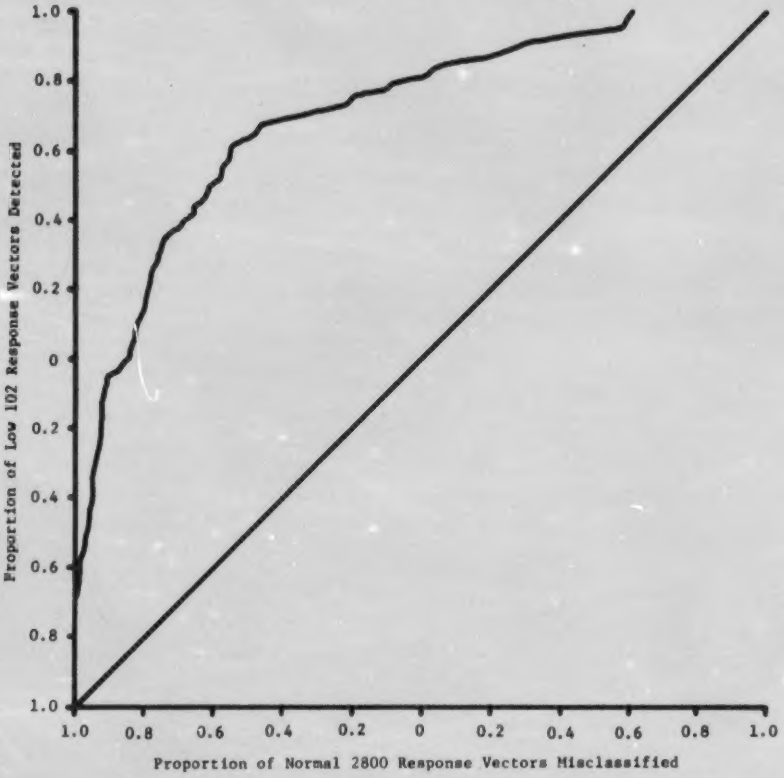
Figure 2
Bivariate Scatterplot of L_0 Computed from Exact and Estimated (from NORMAL 2800 Data) Item Parameters for the First 100 Simulated Examinees in NORMAL 2800 Data



A possible artifact complicating the interpretation of these results is evidenced by consideration of an analogy from multiple regression. The ROC will be high to the degree that the estimated parameters fit NORMAL 2800 better than LOW 102. NORMAL 2800 can be considered analogous to a multiple regression derivation sample and LOW 102 to a cross-validation sample. At least with small samples, overfitting is expected in the derivation sample; and shrinkage, in the cross-validation sample. It may be for this reason that L_0, as a measure of goodness of fit, tends to be smaller in LOW 102. That estimated parameter aberrance scores for the most aberrant LOW 102 examinees are lower than exact parameter aberrance scores supports the suspicion of overfitting. However, any overfitting should result in relatively high NORMAL 2800 scores (i.e., many points above the diagonal in Figure 2) and this was not found. Furthermore, the discrepancy between estimated and exact parameter index values is so small that abcissa values on the ROC would be changed less than .0004 if the lowest LOW 102 index values fell exactly on the diagonal in Figure 1.

If overfitting were a significant artifact, then poor detection would be

Figure 3
ROC Describing Detection of LOW 102 Response Vectors
by L_0 Index Computed from Item Parameters Estimated in NORMAL 2800 Data



expected (1) if the normal group used to evaluate an index was distinct from the norming group utilized for item parameter estimation or (2) if the aberrant and normal groups were pooled to form the norming sample. In Study 4 and in Drasgow (1978) the norming and normal groups were distinct. In the next study the aberrant and normal groups were combined to form a single group used to estimate item parameters.

Study 2: Heterogeneous Norming Sample--Classification and Norming Sample Equal

Problem

In a simulation study all aberrant answer sheets are clearly identified, since they are generated by the experimenter. In an actual study some undetected aberrants are likely to be included in the test norming sample. Will unsuspected aberrants in the norming sample seriously degrade item parameter estimates and undermine the person measurement stage of appropriateness measurement?

A secondary problem is related to the possible overfitting and shrinkage noted in Study 1. Individual simulation examinees are expendable. They can be used for test norming and ignored thereafter because new statistically equivalent answer sheets can easily be generated for use in the person measurement stage. However, in actual studies sample sizes are fixed, and it generally will be important to evaluate appropriateness for the examinees in the norming sample. Will current estimation procedures overfit aberrant examinees in the norming sample, or will it be possible to use an appropriateness index to identify norming sample aberrants?

Method

NORMAL 2800 and LOW 200 were merged to form a data file with a large proportion of aberrant examinees. Item parameters were estimated using all 3,000 simulated examinees. As before, L_0 was computed for all examinees by evaluating the likelihood function at the LOGIST maximum likelihood estimate of ability. New index values for LOW 102 were compared with exact parameter index values, and the L_0 ROC was computed to evaluate detectability.

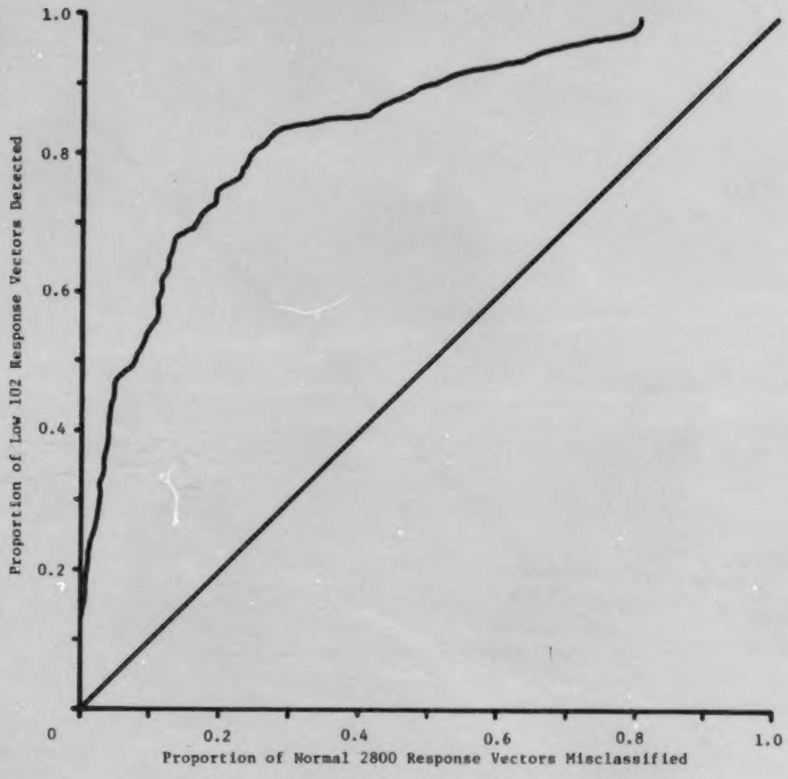
Results and Discussion

Figure 4 shows that estimating item parameters in a large sample with a large proportion of spuriously low examinees need not noticeably degrade appropriateness measurement. The simulation parameter L_0 ROC had hit rates of 11.7%, 21.6%, and 47.1% at false alarm rates of .1%, .9%, and 4.8%. The corresponding heterogeneous norming sample false alarm rates were .1%, 1.2%, and 4.9%. Clearly, the net effect on appropriateness measurement of estimating item parameters from this heterogeneous sample is negligible.

Figure 5 contains the bivariate scatterplot of exact parameter L_0 values plotted against L_0 values computed from item parameters estimated in the heterogeneous sample. The relatively high frequency of points above the diagonal in Figure 5 indicates an overfitting effect for aberrant examinees; however, both Figures 4 and 5 support the conclusion that the effect is small. Figure 5 does so because all points are tightly clustered about the diagonal, i.e., there is little difference between exact parameter L_0 values and heterogeneous sample estimated parameter L_0 values. The high detectability exhibited in the ROC supports the contention that overfitting is small, because a large overfitting effect would have reduced normal-aberrant group differences and therefore reduced detectability of aberrance.

These results on overfitting should be interpreted cautiously. Parameters were estimated in a very large sample (N = 3,000). Further, the nature of spuriously low aberrance may be essential to the small effect. The distinction between bias and sampling error is useful in understanding this point. The process hypothesized to underlie spuriously low aberrance is essentially unsystematic; that is, atypical schooling, alignment errors, and exceptional creativity lead to incorrect responses on different examination items. Thus, the different examinees will tend to have competing effects on item parameter estimates. Consequently, the presence of aberrance in the norming group should affect the sampling error of item parameter estimates to some extent but should have a rela-

Figure 4
ROC Describing Detection of LOW 102 Response Vectors
by the L_0 Index Computed from Item Parameters Estimated in
NORMAL 2800 and LOW 200 Data



tively small effect on the bias of the estimates. In addition, the effect on the sampling error is tolerable due to the large sample size.

Study 3: Actual SAT-V Data-- Overlapping Norming and Classification Sample

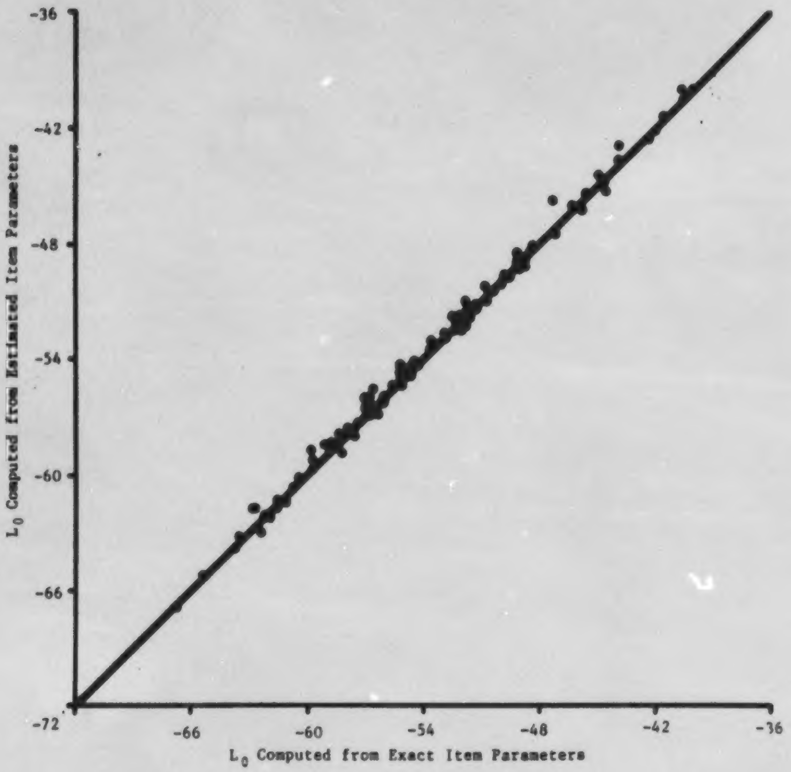
Problem

The one-dimensional 3-parameter logistic model is the most general model for which there is a well-developed parameter estimation literature. It is easy to formulate more plausible refinements and generalizations of this model. However, their use would require a long and costly research program to develop and validate parameter estimation methods. Is the logistic model sufficiently descriptive to detect spuriously low examinees in actual test data?

Method

Three thousand "low omitting rate" examinees were sampled from an administration of the SAT-V. All 3,000 examinees responded to at least 90% of the

Figure 5
Bivariate Scatterplot of L_0 computed from Exact and
Estimated (from NORMAL 2800 and LOW 200 Data) Item Parameters
for LOW 102 Response Vectors



test items. LOGIST was used to estimate item parameters from these 3,000 nominally normal examinees. A file of 200 aberrant examinees was then created by applying the 20% spuriously low modification to answer sheets from examinees 2,801, 2,802,...,3,000.

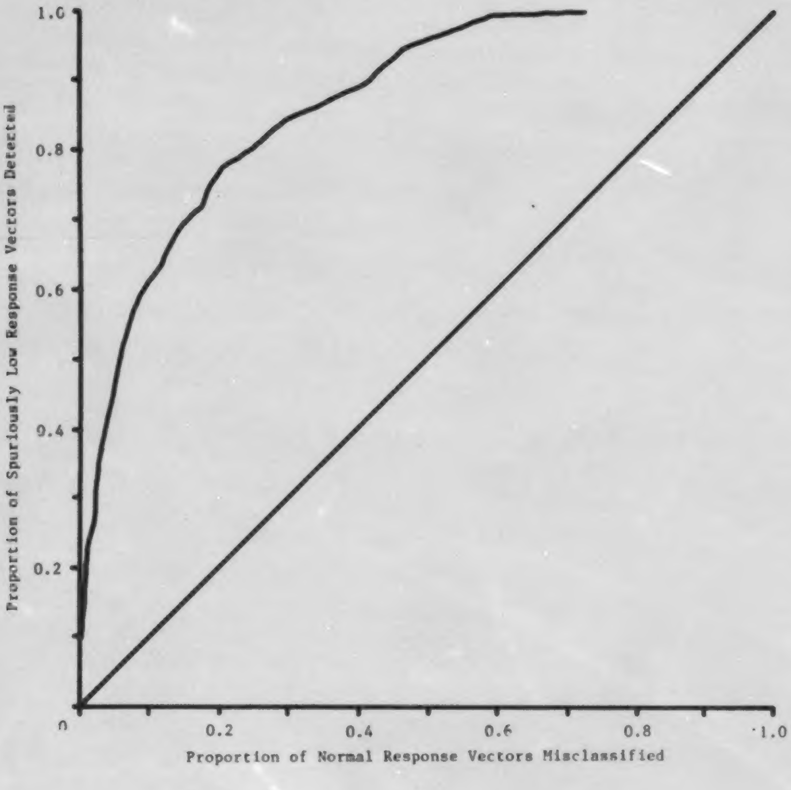
L_0 was computed by maximizing the individual likelihood functions for the 3-parameter logistic model. In this calculation the LOGIST-estimated item parameters were held constant and the likelihood of the individual's response vector was maximized by selecting an optimal ability estimate. The ability estimates are slightly different from the LOGIST ability estimates because the programs used in this study ignored both omitted and not reached items. (The LOGIST procedure ignores not reached items but attempts to use the omitted items by a technique that can be thought of as giving partial credit for omitted items.)

The person parameters of the Gaussian model were estimated by a version of the Fletcher-Powell algorithm (Gruvaeus & Joreskog, 1970). In this calculation each examinee's response vector is considered in isolation; and the likelihood function is maximized to estimate central ability, θ_0, and ability variance, σ^2. As before, omitted and not reached items are ignored.

Results

Figures 6 and 7 present the ROCs for the L_0 and LR appropriateness indices, respectively. It is apparent that high detection rates are obtained at low false alarm rates. In particular, there are .2, .8, and 3.3% false alarms at hit rates of 10%, 20%, and 40% for the L_0 index; and the LR index yields .04%, .8%, and 4.3% false alarms at the same hit rates. These results are even more impressive when it is noted that the aberrant group consists of all 200 spuriously low examinees; no spuriously low examinee was deleted from the analysis due to an insufficient change in item scoring. In fact, 42 examinees had seven or fewer item responses changed (from correct to incorrect and from incorrect to correct) when subjected to the spuriously low modification.

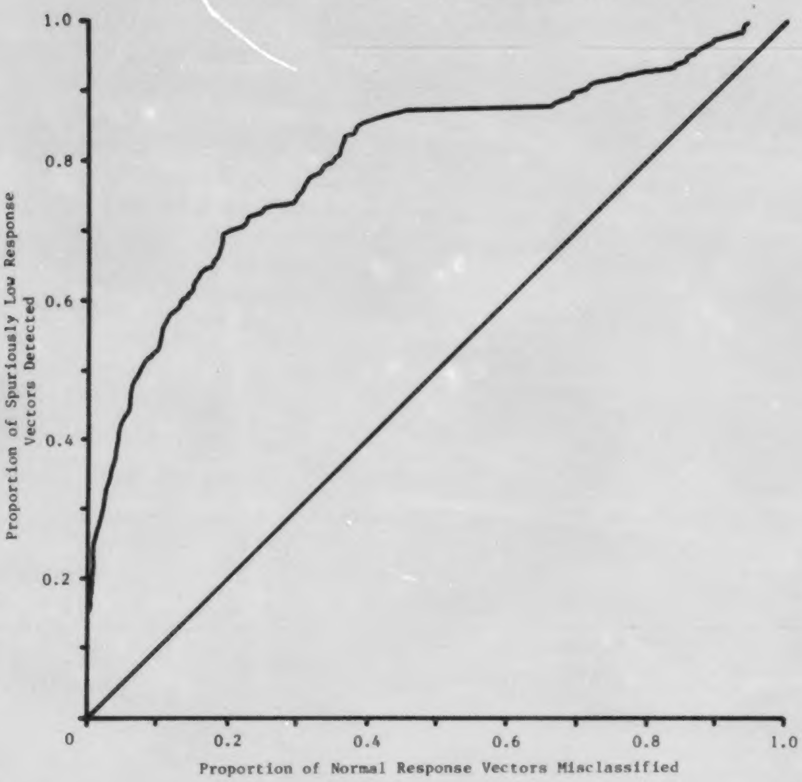
Figure 6
ROC Describing Detection of Spuriously Low Response Vectors by L_0 Index for Actual SAT-V Data



Discussion

It might be argued that the ROC in Figures 6 and 7 capitalize on statistically overfitted data from the normal examinees. This argument rests on the fact that the 2,800 examinees constituting the normal sample were included in the norming sample, whereas item responses from the aberrant examinees were included in the norming sample prior to the spuriously low modification. That is, the post-tampering response vectors were not included in the norming sample for the spuriously low group. It is suspected that the overfitting problem is not serious for two reasons. First, as seen in Studies 1 and 2, overfitting did not create serious difficulties for appropriateness measurement with spurious lowness and for item parameters that are estimated in a large sample. Second, the majority of the item responses made by spuriously low examinees did in fact contribute to item parameter estimation: No more than 20 of the 85 items were modified. Consequently, overfitting had very little, if any, effect on the detection of aberrance in the actual data. However, it seemed wise to attack the overfitting problem directly in Study 4 by separating the test norming and index-evaluation classification samples.

Figure 7
ROC Describing Detection of Spuriously Low Response Vectors by LR Index for Actual SAT-V Data



It is interesting to compare the effectiveness of the L_0 and LR indices in detecting aberrance. Clearly, in Figures 6 and 7 the L_0 index is superior to the LR index at high false alarm rates; but it is expected that there are many more normal examinees than aberrant examinees. Hence, the base rate difference causes the discrepancy to be unimportant between the L_0 and LR ROCs at moderate to high false alarm rates; index performance is crucial only at very low false alarm rates. The LR index detected 19 aberrant examinees (out of 200) without misclassifying a single normal examinee, whereas L_0 detected only 6 without misclassification. At a false alarm rate of .5%, LR detected 37 aberrant examinees and L_0 detected 32. Thus, at very low false alarm rates the LR index seems somewhat more powerful.

Study 4: Actual GRE-V Data-- Distinct Norming and Classification Samples

Problem

The positive results in Study 3 may be attributable in large part to overfitting made possible by overlapping norming and classification samples. In this study the samples were independent. In addition, Study 4 investigated the generality of the appropriateness indices. The indices used in Studies 1 through 3 were selected by Levine and Rubin (in press) from a larger collection of indices because of their superior performance in experimental studies with actual and simulated SAT-V data. The question is, are these methods applicable to other tests?

Method

The responses to the Verbal Section of the Graduate Record Examination (GRE-V) by 10,000 examinees were utilized in the following manner. First, a file of 3,000 examinees (FILE1) with a wide range of ability and unrestricted omitting was created by selecting Examinees 1, 2, 3, 11, 12, 13, ..., 9991, 9992, 9993. This data set was then analyzed by LOGIST to obtain item parameter estimates. A second file was created by examining the item responses of the remaining 7,000 examinees and selecting examinees with a low omitting rate. A total of 2,470 examinees who had answered at least 86 of the 95 GRE-V items was obtained. Two hundred of these examinees were subjected to the 20% spuriously low modification. These modified response vectors formed the aberrant group for Study 4, and the remaining 2,270 response vectors formed the normal group.

L_0 was computed for the 200 aberrant and 2,270 normal response vectors as in Study 3 using the FILE1 item parameter estimates. Notice that here the test norming sample was distinct from the normal and aberrant samples.

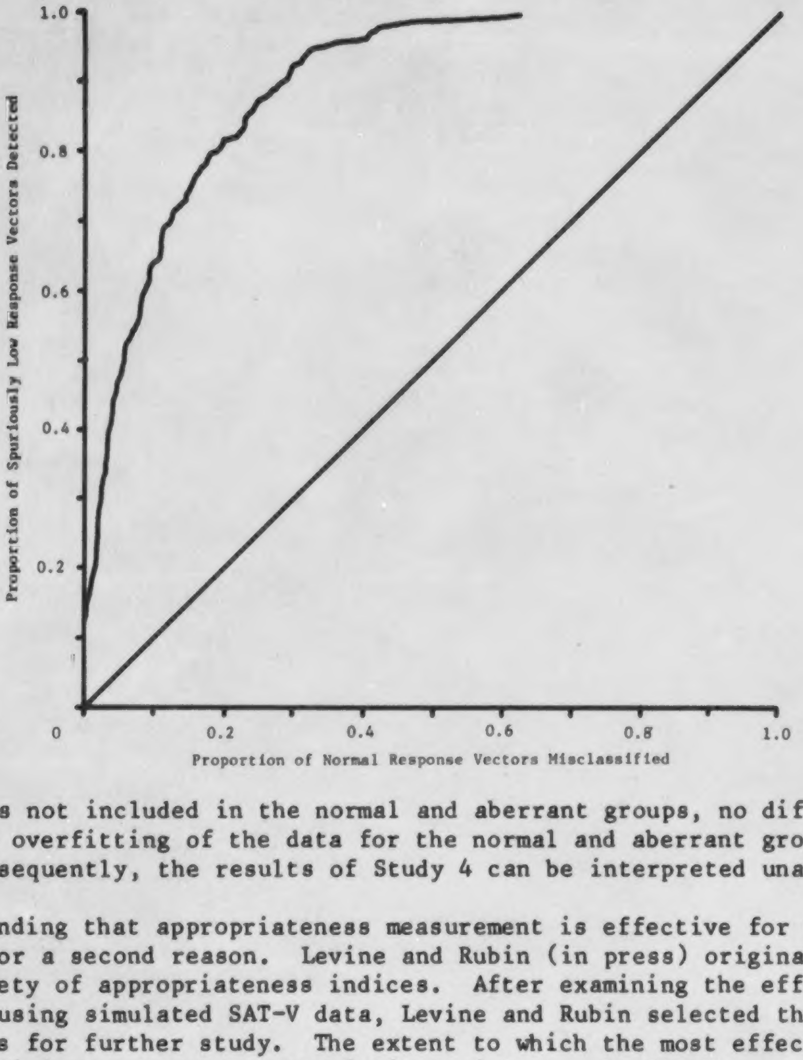
Results

Figure 8 presents the ROC for the L_0 index. Clearly, detection rates are quite high. At hit rates of 10%, 20%, and 40% there are .3%, 1.4%, and 3.7% false alarms. These detection rates were not substantially different from those obtained in Study 3.

Discussion

The results of Study 4 are important for two reasons. First, the criticism of overfitting the normal sample, which could be made for Studies 1 and 3, is not relevant for Study 4. Because the test norming sample (FILE1) was composed

Figure 8
ROC Describing Detection of Spuriously Low Response Vectors by L_0 Index for Actual GRE-V Data



of examinees not included in the normal and aberrant groups, no differential statistical overfitting of the data for the normal and aberrant groups was possible. Consequently, the results of Study 4 can be interpreted unambiguously.

The finding that appropriateness measurement is effective for GRE-V data is important for a second reason. Levine and Rubin (in press) originally considered a variety of appropriateness indices. After examining the effectiveness of each index using simulated SAT-V data, Levine and Rubin selected the most effective indices for further study. The extent to which the most effective indices were capitalizing on test characteristics unique to the SAT-V was unknown. Study 4 shows that the methods developed using SAT-V data were sufficiently general to be implemented for GRE-V data.

Summary and Conclusion

These studies have shown that some appropriateness measurement techniques are robust to errors in estimation of item parameters, to the inclusion of unidentified aberrants in the test norming sample, and to violations of the 3-parameter logistic model, which surely must exist in actual data. The detection rate of spuriously low examinees was high in all the analyses undertaken.

Model Validity and Variable Ability Models

The 3-parameter logistic model, with its local independence and unidimensionality assumptions, is admittedly simplistic. Lord (1975b) has brought to the attention of the authors a very likely violation of local independence on the SAT-V and GRE-V. Some paragraphs associated with several items on the reading comprehension part of the examinations may be misunderstood or, alternatively, may be relevant to an area in which the examinee has an unusually strong background. Thus, it seems virtually certain that responses to items referring to the same passage will be more highly interrelated than the local independence assumption (Equation 1) predicts. However, in spite of its shortcomings, the standard model has been able to describe regularities in data well enough to support appropriateness measurement. A more valid model, presumably, could support even more powerful appropriateness indices.

Parenthetically, it is noted that violations of local independence can be accommodated by variable ability generalizations of the standard latent trait model. A variable ability model is now being considered for dealing with covarying blocks of items (such as those called to the authors' attention by Lord, 1975b), the blocks model, in which a test is analyzed into interrelated blocks of items. The examinee's ability on an item in a particular block is his or her central ability, θ_0, plus a (normally) distributed correction. This correction is constant throughout the block of items. The Gaussian model is a limiting special case in which each item forms a one-item block. The standard model is another limiting case in which the entire test forms one block. If an adequate item parameter estimation procedure were developed for the blocks model, a substantial improvement in appropriateness measurement could be achieved.

The variable ability models are related to the independent work of Lumsden (1978). Lumsden used "person characteristic curves" (PCCs) to describe fluctuations in ability. The authors' work with the Gaussian model appears to be a quantitative step in the direction Lumsden recommends for his purposes.

Item Parameters versus Conditional Probabilities

In appropriateness measurement and in many other applications of latent trait models, even very large standard errors of item estimates can sometimes be tolerated. This is true because the probability measure determined by a latent trait model depends directly on the conditional probability functions or item characteristic curves (ICCs), P_i, and only indirectly on the item parameters. The item parameters are simply a convenient way to encode the shape of ICCs. In fact, some curves are adequatey described by a broad range of parameters.

This point is developed in a recent study of item bias (Linn, Levine, Hastings, & Wardrop, 1979). In that study, item parameters were estimated for a reading achievement test from two groups with widely different distributions of achievement scores. A remarkable degree of invariance was observed when the estimated curves from the two groups were compared. The two sets of estimated curves were generally very nearly the same, although a superficial comparison of parameters often showed large differences in the estimated item parameters.

Estimation of ICCs Having More Than One Parameter

In spite of several monte carlo studies and numerous successful applications of the 3-parameter logistic model, there seems to be some doubt about whether the 3-parameter ICCs can be estimated by currently used programs or by any method whatsoever. Some psychometricians evidently believe that adequate parameter estimates can be obtained only with the 1-parameter, or Rasch, model (i.e., the specialization of the 3-parameter model obtained by setting $a_i = 1$ and $c_i = 0$ in Equation 3).

In fact, parameter estimation techniques are available for models with much more complex ICC shapes than that of the 3-parameter logistic model. Lord (1970) and Samejima (1977) have formulated, programmed, and applied nonparametric curve estimation techniques suitable for estimating curves of arbitrary shape. Furthermore, Levine (1976) has proved a consistency result for an estimation technique for estimating points on curves of arbitrary shape. It seems that the estimation difficulties ensuing from the departure from very simple curve shapes have been exaggerated.

Work in Progress

The next step in the development of appropriateness measurement will be to develop techniques for conventional tests in which there is substantial omitting. The research in this paper with actual data has been restricted to answer sheets with 90% or higher response rates. However, a substantial proportion of the examinees have omitting rates greater than 10% on the SAT-V and GRE-V.

It has been found that there is an orderly relationship between omitting and ability on many items (Levine, Drasgow, & Rubin, in prep.) on conventional tests; and latent trait models for polychotomously scored items are being developed to exploit this relationship. By treating "omitted" and "not reached" as option choices, each answer sheet can be analyzed as if every item were answered. The preliminary investigations indicate that in this way both the range of applicability and the statistical power of appropriateness measurement can be significantly increased.

REFERENCES

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Donlon, T. F., & Fischer, F. E. An index of an individual's agreement with group determined item difficulties. *Educational and, Psychological Measurement*, 1968, *28*, 105-113.

Drasgow, F. Statistical indices of the appropriateness of aptitude test scores (Doctoral dissertation, University of Illinois, 1978). *Dissertation Abstracts International*, 1979, *39*, 6095B. (University Microfilms No. 79-13435)

Flier, H. van der. Environmental factors and deviant response patterns. In Y. H. Portinga (Ed.), *Basic problems in cross-cultural pyschology*. Amsterdam: Swets & Zeitlinger, 1977.

Ghiselli, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, 1956, *40*, 374-377.

Ghiselli, E. E. Differentiation of tests in terms of the accuracy with which they predict for given individual. *Educational and Psychological Measurement*, 1960, *20*, 675-684. (a)

Ghiselli, E. E. The prediction of predictability. *Educational and Psychological Measurement*, 1960, *20*, 3-8. (b)

Gruvaeus, G. T., & Jöreskog, K. G. *A computer program for minimizing a function of several variables* (Research Bulletin 70-14). Princeton, NJ: Educational Testing Service, 1970.

Hambleton, R.K., & Rovenelli, R.A. A FORTRAN IV program for generating response data from logistic test models. *Behavioral Sciences*, 1973, *18*, 74.

Levine, M. V. *Item-item curves and consistent mental test parameter estimates* (Research Bulletin 76-36). Princeton, NJ: Educational Testing Service, 1976.

Levine, M.V., Drasgow, F., & Rubin, D. B. *On the relationship between ability and probability of option selection*, in preparation.

Levine, M. V., & Rubin, D. B. Measuring the appropriateness of multiple-choice test scores. *Journal of Eductional Statistics*, in press.

Levine, M. V., & Saxe, D. H. *The use of periodic functions to measure the difficulty of aptitude test items* (Research Bulletin 76-17). Princeton, NJ: Educational Testing Service, 1976.

Linn, R. L., Levine, M. V., Wardrop, J. L., & Hastings, C. N. *An investigation of item bias in a test of reading comprehension*, in preparation.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F. M. Item characteristic curves estimated without knowledge of their mathematical form--a confrontation of Birnbaum's logistic model. *Psychometrika*, 1970, *35*, 43-50.

Lord, F. M. *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters* (Research Bulletin 75-33). Princeton, NJ: Educational Testing Service, 1975. (a)

Lord, F. M. Personal communication, 1975. (b)

Lumsden, J. Tests are perfectly reliable. British Journal of Mathematical and Statistical Psychology, 1978, 31, 19-26.

Samejima, F. A method of estimating item characteristic functions using the maximum likelihood estimate of ability. Psychometrika, 1977, 42, 163-191.

Wood, R. L., & Lord, F. M. A user's guide to LOGIST (Research Bulletin 76-4). Princeton, NJ: Educational Testing Service, 1976.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Bulletin 76-6). Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problms with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

ACKNOWLEDGMENTS

This research was primarily supported by funds provided by the Educational Testing Service (ETS) and grants from the Graduate Record Examination Board (No. 75-3) and the College Entrance Examination Board (PJ 82201) to Michael V. Levine. The points of view and opinions expressed here do not represent the position or policy of ETS, GREB, or CEEB.

The research reported here was initiated by the first author while he was at ETS. Much of this article was written while the second author was at the University of Illinois.

DISCUSSION: SESSION 7

JAMES LUMSDEN
UNIVERSITY OF WESTERN AUSTRALIA

The three papers--by Mead, by Wainer and Wright, and by Levine and Drasgow--agree that the task of estimating ability is a matter of estimating a parameter of the person characteristic curve (PCC), which was first suggested by Mosier (1940, 1941). It was rediscovered independently by Weiss (1973) and by me (Lumsden, 1976), but not independently because I had read the Mosier paper and had "forgotten" about it. The PCC is the plot of proportion passed against item difficulty for a single individual.

Each of the papers assumes--but Levine and Drasgow's not quite so completely--that all we need do (in a proper world) is to estimate the location parameters. All the PCCs "ought" to have the same slope; and departures from the ideal represent aberrances, perturbations, warts on the face of the estimate of a person's ability.

Mead has proposed to apply vanishing cream to the warts, at least to the warts he sees. He has eliminated responses considered aberrant and has estimated ability from those remaining. Wainer and Wright have made some comments on the inefficiency of this pruning procedure and have suggested a rather more suitable one. What I am concerned about is the suggestion that if the Mead procedure (or the similar dropping of aberrant subjects) is carried out, the Rasch model has been fitted. In no sense has this been done.

What about the possibility of a Type 2 error? Consider the following simple, but quite plausible, example. Two people each know the answer to only one item of a seven-item test. They are each somewhat lucky and guess the answer to three others. They produce the following response vectors, with items ordered in difficulty from left to right:

A 1 0 0 0 1 1 1
B 1 1 1 1 0 0 0

The Mead procedure would eliminate for Subject A the final three correct answers and score only the first correct. For Subject B the score would be estimated as if B knew the answer to four items. I submit that each of these subjects equally fits or does not fit the Rasch model.

The suggestion that a model can be fitted by removing the data that do not fit it is nonsense, dangerous nonsense. Brown and Stephenson (1932) did this when they claimed a fit for the Spearman two-factor theory. Wolfle (1940) commented, "if you remove all the variables that do not meet the tetrad-difference criterion, those that are left do meet it."

Wainer and Wright have proposed to shave their subjects with a jackknife and thus clip the warts. The general proposal I find attractive. It gives less weight to outliers but does not eliminate them: All the subjects and all the results are considered. In this it is rather like the work of the psychophysicists who use the Mueller-Urban weights.

Wainer and Wright's results should be taken with a square root of salt. When someone talks of estimating a point value such as a person's ability, I think of confidence limits and standard errors rather than variances. If the square root of all of Wainer and Wright's results are taken, the differences between methods are seen as very much less; and in some cases, they can be described as negligible. It should be noted, too, that the method of reporting only efficiency ratios suppresses the main effect of test length. What is the value of all this arithmetic? One item, or two or three?

I would also like to see the bias and precision estimates reported separately. It should be noted that the efficiency comparisons are critically dependent on the selection of the a value for the items. If the a value is increased, the efficiency differences will be less.

Wainer and Wright's treatment of omitted responses is distressingly foolish. They permitted subjects to omit items and did not punish them in any way but simply scored the other items as if they comprised the entire test. Now, if there are any person-item interactions, the person smart enough (or lazy enough) to omit those items whose answer does not come to him/her immediately will be given a higher score than the person who attempts all questions. This is a seriously biasing part of Wainer and Wright's procedure. One can usually be certain that a person who omits does not know the correct answer. He/she should either be counted as wrong or should be given the chance expectation of being correct.

Both the Mead paper and the Wainer and Wright paper approach the problem of fitting the Rasch model with what I term the psychoarithmetician's fallacy: that arithmetic can be substituted for experimental control. Why not attempt to meet the strict requirements of the Rasch model? One could begin by eliminating the problem of guessing by using only completion-type items. If this means that some tests have to be hand-scored, so much the better. One should then attempt to construct a strictly unidimensional test, keeping the test construction completely independent of the test scoring and application phases.

Levine and Drasgow used the 3-parameter model, also seeming to believe that arithmetic is preferable to the "restrictions" of experimental control and agreeing that perturbations of the PCC are aberrances, warts. Shuddering, they dismissed the warty ones from consideration. Some of the things they chose to call aberrant seem strange to me: If a test requires a subject to be able to read and to understand English, in what way is it aberrant when it gives a low score to someone who cannot read and understand English? The positive feature of Levine and Drasgow's paper is that they do consider Type 2 errors and that their ROC curve procedure is ingenious and generalizable to a variety of other situations.

The assumption underlying all three papers--that the slopes of the PCC "ought" to be the same--is simply that: an assumption. If it is agreed that it is plausible that people do have fluctuations in their ability, then it seems not implausible to assume that people may differ reliably and significantly in the extent of this fluctuation. There is evidence, admittedly not overwhelming, from Mosier (1940, 1941) and Weiss (1973) that they do. Further evidence should be sought. If there are reliable differences in the slopes, it is difficult to see how these differences can be distinguished from the aberrances that have so distressed our participants.

REFERENCES

Brown, N., & Stephenson, W. A test of the theory of two factors. British Journal of Psychology, 1933, 23, 352-370.

Lumsden, J. Test theory. Annual Review of Psychology, 1976, 27, 251-280.

Mosier, C. I. Psychophysics and mental test theory: Fundamental postulates and elementary theorems. Psychological Review, 1940, 47, 355-366.

Mosier, C. I. Psychophysics and mental test theory II. The constant process. Psychological Review, 1941, 48, 235-249.

Weiss, D. J. The stratified adaptive computerized test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

Wolfle, D. Factor analyses to 1940. Psychometric Monographs, 1940 (No. 3).

SESSION 8:
USE OF LATENT TRAIT MODELS WITH ESTIMATED ITEM PARAMETERS

ROBUSTNESS OF LATENT TRAIT MODELS AND EFFECTS OF TEST LENGTH AND SAMPLE SIZE ON THE PRECISION OF ABILITY ESTIMATES

RONALD K. HAMBLETON
AND LINDA L. COOK
UNIVERSITY OF MASSACHUSETTS

ESTIMATING ABILITIES WITHIN THE TWO-PARAMETER LOGISTIC LATENT TRAIT MODEL IN THE PRESENCE OF A NON-SYMMETRIC DISTRIBUTION OF ABILITY

MICHAEL WALLER
UNIVERSITY OF WISCONSIN--
MILWAUKEE

ESTIMATION OF PARAMETERS IN THE 3-PARAMETER LATENT TRAIT MODEL

HARIHARAN SWAMINATHAN
AND JANICE GIFFORD
UNIVERSITY OF MASSACHUSETTS

SMALL N JUSTIFIES RASCH METHODS

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

DISCUSSION

BERT F. GREEN, JR.
JOHNS HOPKINS UNIVERSITY

THE ROBUSTNESS OF LATENT TRAIT MODELS AND EFFECTS OF TEST LENGTH AND SAMPLE SIZE ON THE PRECISION OF ABILITY ESTIMATES

RONALD K. HAMBLETON AND LINDA L. COOK
UNIVERSITY OF MASSACHUSETTS

Although latent trait models are potentially very useful, there remain many practical problems at the application stage. For example, how should a latent trait model be selected? It is tempting to use the more general models, since these models will provide the "best" fits to the available test data. Unfortunately, the more general latent trait models (for example, the 3-parameter logistic test model) require more computer time to obtain satisfactory solutions, larger samples of examinees, and longer tests and are more difficult for practitioners to work with. Clearly, more needs to be known about the goodness of fit and robustness of latent trait models. Such information would aid practitioners in the important step of selecting a test model.

There has been some research on the goodness of fit of different latent trait models to a variety of test data sets (e.g., Lord, 1975; Tinsely & Dawis, 1977; Wright, 1968), and generally the results have been good (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978). However, only one study compared the fit of more than one latent trait model to the same data sets (Hambleton & Traub, 1973). In that study, improvements were obtained in predicting test score distributions (for three tests) from the 2-parameter model as compared to the 1-parameter model.

On the question of model robustness (i.e., the extent to which the assumptions underlying the test model can be violated to a greater or lesser extent by the test data and be fitted by the model), the results of several studies have been reported (Dinero & Haertel, 1977; Hambleton, 1969; Hambleton & Traub, 1976; Panchapakesan, 1969). These results have been mixed, perhaps because of the confounding of results with sample sizes.

The problem with most of the goodness-of-fit studies and the robustness studies reported to date is that they provide no indication of the practical consequences of fitting a less than perfect model to a data set. It really is of little interest to the practitioner to know that 15 out of 20 items failed to be fitted by a test model when the range of discrimination parameters reached a value of .80. If the size of the examinee sample is large enough, probably all items could be identified by a chi-square statistic of goodness of fit as not fitting the model. On the other hand, if the size of the examinee sample is small enough, perhaps none of the items would be misfit by the model. Study 1 addressed this question.

One of the features of using any latent trait model is the possibility of specifying a "target information curve" and then selecting test items from an item pool to produce a test with the features characterized by that curve. A target information curve describes the desired level of information at each point on the ability scale underlying examinee test performance. Information, in turn, is directly related to the degree of precision of ability estimates at different points on the ability continuum. In fact, as long as a test is not too short, the standard error of estimate at a particular ability level is equal to 1 divided by the square root of information provided by the test at the ability level in question. In practice, since the contribution of each test item to the test information curve (referred to as a "score information curve" when item parameter estimates are used instead of the item parameter values) is known--that is, once the item parameter values or the item parameter estimates are specified--it is possible to select test items from a pool of calibrated test items (i.e., a pool of test items with associated parameter estimates) to produce a score information curve that approximates a desired target information curve.

One of the problems with the paradigm offered above for test development is the imprecision associated with the item parameter estimates. Score information curves--and therefore the associated standard errors of ability estimates--will depend on the precision of item parameter estimates. In turn, precision of item parameter estimates is influenced by the examinee sample size used to estimate the item parameters, and in the case of the item discrimination parameter, estimates are influenced by the length of the test. Study 2 was designed to address this issue.

STUDY 1

The purpose of Study 1 was to study systematically the goodness of fit of the 1-, 2-, and 3-parameter logistic models. Using computer-simulated test data, the effects of four variables were studied: (1) variation in item discrimination parameters, (2) the average value of the pseudo-chance-level parameters, (3) test length, and (4) the shape of the ability distribution. Artificial or simulated data representing departures of varying degrees from the assumptions of the 3-parameter logistic test model were generated and the goodness of fit of the three test models to the data were studied.

Method

Simulating the Test Data

The simulation of item response data for examinees was accomplished using the 3-parameter logistic model. First, the number of examinees (N), shape of the ability distribution, and values of the ability parameters (θ_i = 1, 2,...,N) were specified. Next, the number of items in the test ($\underline{n}$) and values of the three item parameters ($\underline{a}_g$, $\underline{b}_g$, $\underline{c}_g$, $\underline{g}$ = 1, 2,...,$\underline{n}$) were specified. Then, the examinee and item parameters were substituted in the equation of the 3-parameter logistic model to obtain $\underline{p}_{ij}$ ($0 \leq \underline{p}_{ij} \leq 1$), representing the probability that examinee $\underset{\sim}{i}$ correctly answered item $\underline{j}$. The probabilities were arranged in a matrix $\underset{\sim}{P}$ of order N × $\underline{n}$ whose $(\underline{i}, \underline{j})^{th}$ element was $\underline{p}_{ij}$. $\underset{\sim}{P}$ was then converted into a matrix of the item scores for examinees (1 = correct answer, 0 = incorrect answer) by comparing each $\underline{p}_{ij}$ with a random number obtained from a uniform distribution in the interval 0 to 1. If the random number was less than or equal to $\underline{p}_{ij}$ (which would happen on the average $\underline{p}_{ij}$ of the time), $\underline{p}_{ij}$ was set equal to 1; otherwise, $\underline{p}_{ij}$ was set to 0. The matrix $\underset{\sim}{P}$ of 0's and 1's was the simulated test data. Three statistics used in estimating examinee ability were calculated:

1-parameter score, $\sum_{g=1}^{n} u_g$, the number-correct score;

2-parameter score, $\sum_{g=1}^{n} a_g u_g$, and

3-parameter score, $\sum_{g=1}^{n} w_g(\theta)\, u_g$,

corresponding to statistics that are used in the estimation of examinee ability with the 1-, 2-, and 3-parameter models, respectively. For the 3-parameter model statistic, since the item weights $[\underline{w}_g(\theta)]$ depend on examinee ability, 3-parameter model estimates of ability were estimated for each examinee from LOGIST (Wood, Wingersky, & Lord, 1976).

The values of item parameters used are summarized in Table 1.

Item parameters. Two test lengths (20 and 40 items) were used in the simulations. Item difficulty parameters, $\underline{b}$, were selected at random from a uniform distribution in the interval [-2,2]. An analysis of the difficulty parameters reported by Lord (1968) suggested that this decision was reasonable.

The discrimination parameters, $\underline{a}$, were selected at random from a uniform distribution with mean = 1.12. The range of the discrimination parameters was a variable under investigation. The range was varied from 0.0 to a maximum of 1.24 [.50 to 1.74], and an intermediate value of .62 [.81 to 1.43] was also studied. The maximum value of discrimination was similar to the range and distribution of the discrimination parameters reported for the Verbal Section of the Scholastic Aptitude Test (SAT-V; Lord, 1968).

The extent of guessing in the simulated test data was another variable under study. Two values of the average guessing parameter were considered: $\underline{c}$ = 0.0 and $\underline{c}$ = .25. All pseudo-chance-level parameters were set equal to the mean value of the $\underline{c}$ parameter under investigation.

Examinee parameters. The number of examinees was set to 500. This number was sufficient to produce stable goodness-of-fit results. Two distributions of ability were considered: Uniform [-2.5, 2.5] and Normal [0, 1].

Table 1
Test Lengths, Range of Discrimination Parameters, and Pseudo-Chance Level Parameters for Each Data Set

Data Set	Test Length	Variation in Discrimination Parameters	Pseudo-Chance Level Parameters
A	20	0.00	.00
B	20	0.00	.25
C	20	.81 to 1.43	.00
D	20	.81 to 1.43	.25
E	20	.50 to 1.74	.00
F	20	.50 to 1.74	.25
G	40	0.00	.00
H	40	0.00	.25
I	40	.81 to 1.43	.00
J	40	.81 to 1.43	.25
K	40	.50 to 1.74	.00
L	40	.50 to 1.74	.25

Goodness of Fit

For each data set A through L—2 test lengths × 2 levels of pseudo-chance parameters × 3 levels of variation in discrimination parameters--and for each of the two ability distributions--Uniform and Normal--three scoring methods were used to estimate ability based on the 1-, 2-, and 3-parameter models. Since simulated data were used, it was possible to "know" examinee ability scores, which served as the criterion against which to judge the statistics derived from the three test models for ranking examinees. The rankings of examinees derived from each model (for each set of test data) were then compared to examinee "true" abilities using Spearman rank-difference correlations and the average discrepancy in ranks. Because of the arbitrariness of the scale on which $\hat{\theta}$ is measured, summary statistics such as $\sum_{i=1}^{N} |\theta_i - \hat{\theta}_i|/N$ were not studied. To further facilitate the interpretation of results, they are reported separately for each half of the ability distribution as well as for the total ability distribution.

Results

Results are summarized in Tables 2 through 5.

Level of Variation in Discrimination Parameters

For the values studied in the paper, using discrimination parameters as item weights contributed very little to the proper ranking of examinees.

Table 2
Spearman Rank-Order Correlations (r) and
Average Absolute Difference in Rank Orders (AAD) for
the Two Halves of the Uniform Ability Distribution

Data	True vs. 1-P Model		True vs. 2-P Model		True vs. 3-P Model	
Set	r	AAD	r	AAD	r	AAD
Lower Half (θ = -2.5 to 0.0)						
A	.88	54.24	.88	54.24	.88	54.24
B	.77	76.61	.77	76.61	.83	64.98
C	.88	56.07	.88	56.41	.88	56.40
D	.76	77.14	.76	76.90	.83	64.28
E	.87	56.50	.87	56.56	.87	56.56
F	.75	80.08	.75	79.92	.83	65.77
G	.94	36.48	.94	36.48	.94	36.48
H	.87	58.58	.87	58.58	.91	48.70
I	.95	36.50	.95	36.47	.95	36.47
J	.87	57.66	.88	56.86	.91	48.01
K	.94	37.86	.95	36.96	.95	36.74
L	.87	57.82	.88	56.87	.91	48.22
Upper Half (θ = 0.0 to +2.5)						
A	.88	54.45	.88	55.62	.88	55.62
B	.84	63.68	.83	65.35	.83	65.73
C	.89	52.23	.88	55.38	.88	55.38
D	.88	63.80	.83	65.02	.84	63.19
E	.87	56.99	.88	55.38	.88	55.47
F	.80	71.57	.80	70.72	.80	69.16
G	.94	39.03	.94	40.50	.94	40.50
H	.90	50.19	.90	51.05	.90	50.85
I	.94	40.65	.93	41.83	.93	41.85
J	.91	49.14	.90	50.55	.91	50.27
K	.93	40.79	.94	38.93	.94	38.94
L	.89	52.88	.89	52.90	.89	52.68

Level of Pseudo-Chance-Level Parameters

With the 20-item tests the 3-parameter model was considerably more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .08 higher (about .75 to .83) in the uniform distribution of ability and about .08 higher in the normal distribution (about .65 to .73). The improvement in the average absolute difference in rank order was about 13.

With the 40-item tests, the 3-parameter model was also somewhat more effective at ranking examinees correctly in the lower half of the ability distribution. Correlations were about .04 higher in both ability distributions. The improvement in the average absolute difference in rank order was about 8. The reduction in effectiveness of the 3-parameter model weights was to be expected with the longer tests. Gulliksen (1950) noted the insignificance of scoring weights when the test gets longer and test items are positively correlated.

Table 3
Spearman Rank-Order Correlations (r) and
Average Absolute Difference in Rank Order (AAD) for
the Full Uniform Ability Distribution ($\theta = -2.5$ to $+2.5$)

Data Set	True vs. 1-P Model		True vs. 2-P Model		True vs. 3-P Model	
	r	AAD	r	AAD	r	AAD
A	.97	28.26	.97	28.37	.97	28.37
B	.93	41.85	.93	41.97	.95	36.97
C	.97	28.81	.97	29.14	.97	29.14
D	.93	42.40	.93	43.93	.94	38.59
E	.97	30.83	.97	30.14	.97	30.14
F	.93	42.20	.93	42.73	.94	39.02
G	.98	20.44	.98	20.61	.98	20.61
H	.96	30.13	.96	30.26	.97	27.02
I	.98	21.09	.98	21.25	.98	21.25
J	.96	30.69	.96	30.75	.97	27.74
K	.98	22.48	.98	21.81	.98	21.81
L	.96	31.49	.96	30.50	.97	27.30

For examinees in the upper half of the ability distribution, and for the data sets studied, the number-correct score was about as effective as the more complicated scoring weights used in the 2- and 3-parameter models.

Shape of the Ability Distribution

As expected, correlations tended to be higher for the uniformly distributed ability scores.

Test Length

Increases in correlations were observed due to doubling the length of the test. Again, as expected, they tended to be rather small.

Conclusions

From the data in this study, it is clear that there are some sizable gains to be expected in the correct ordering of examinees at the lower end of the ability continuum with modest length tests ($n = 20$) when 3-parameter model estimates are used (as opposed to the number-correct score). The gains were cut roughly in half when the tests were doubled ($n = 40$) in length. It was also surprising that item discrimination parameters as weights had so little effect on the results. However, Gulliksen (1950) summarized the research on item weights nearly 30 years ago and came to essentially the same conclusion. Consequently, to the extent that these simulated data sets are typical of real data, it would appear that the application of latent trait models to the problem of ranking examinees is probably not worth the trouble except in those situations where gains of the size noted for lower ability examinees are important. The number correct score ranks examinees nearly as well as the most complicated scoring methods.

Table 4
Spearman Rank-Order Correlations (r) and
Average Absolute Difference in Rank Order (AAD) for
for the Two Halves of the Normal Ability Distribution

Data	True vs. 1-P Model		True vs. 2-P Model		True vs. 3-P Model	
Set	r	AAD	r	AAD	r	AAD
Lower Half (θ = 0.0, SD_θ = 1.0)						
A	.82	65.58	.82	65.58	.82	65.58
B	.65	94.93	.65	94.93	.74	82.54
C	.84	62.72	.83	63.26	.83	63.31
D	.65	95.18	.65	95.77	.73	83.49
E	.80	70.65	.80	69.43	.80	69.41
F	.66	94.63	.64	95.80	.73	83.38
G	.91	46.03	.91	46.03	.91	46.03
H	.81	68.70	.81	68.70	.85	61.63
I	.90	48.23	.91	47.28	.91	47.28
J	.81	68.08	.82	67.05	.85	60.09
K	.90	48.22	.91	46.58	.91	46.58
L	.81	69.01	.81	68.66	.85	61.58
Upper Half (θ = 0.0, SD_θ = 1.0)						
A	.84	60.51	.84	60.81	.84	60.81
B	.76	75.75	.76	76.16	.77	75.08
C	.85	61.09	.85	61.60	.85	61.61
D	.76	76.41	.76	78.02	.77	75.63
E	.83	64.79	.85	63.08	.85	63.08
F	.75	78.69	.75	79.92	.77	77.01
G	.90	50.71	.90	50.75	.90	50.75
H	.82	65.18	.82	65.45	.83	64.24
I	.89	51.25	.90	50.21	.90	50.23
J	.82	65.92	.83	64.84	.84	63.16
K	.89	51.01	.90	49.95	.90	49.95
L	.81	67.60	.82	64.51	.83	63.96

The results of this single study should be generalized with caution, since the values of the item parameters used may not be typical of real data sets. Secondly, the criterion measure of goodness of fit seems suitable for the situation in which a user desires to make norm-referenced interpretations of test scores. There are many other test situations (for example, those involving adaptive tests, test score equating, and criterion-referenced tests) where a different criterion to judge the quality of a solution would be more suitable. Thirdly, these results provide a somewhat unfair comparison of the 2-parameter model with the other two models because the item discrimination parameters used in the weighting process to derive statistics for ability estimation would have been somewhat different had the "best-fitting" 2-parameter curves to the 3-parameter item characteristic curves been used. The item discrimination parameters in the best fitting 2-parameter curves would have differed somewhat from those defined in the 3-parameter curves to which they were fitted. Finally, the

Table 5
Spearman Rank-Order Correlations (r) and
Average Absolute Difference in Rank Order (AAD) for
the Full Normal Ability Distribution ($\overline{X}_\theta$ = 0.0, SD_θ = 1.0)

Data	True vs. 1-P Model		True vs. 2-P Model		True vs. 3-P Model	
Set	r	AAD	r	AAD	r	AAD
A	.94	36.84	.94	36.91	.94	36.91
B	.88	53.94	.88	53.90	.91	47.55
C	.94	35.87	.94	35.99	.94	35.98
D	.88	54.31	.88	54.34	.91	48.61
E	.93	41.11	.93	40.96	.93	40.96
F	.87	55.73	.87	57.94	.88	53.13
G	.97	26.60	.97	26.62	.97	26.62
H	.95	36.44	.95	36.46	.96	33.03
I	.97	25.20	.97	25.54	.97	25.53
J	.94	38.86	.94	37.65	.95	34.15
K	.97	27.04	.97	25.88	.97	25.87
L	.94	38.79	.94	37.33	.95	34.68

correlation results of the 1-parameter model and, to a much lesser extent, the 2-parameter model are inflated (to an unknown extent) because of tied scores. Therefore, the true differences in the reported correlations are somewhat larger than those reported in Tables 1 to 5.

STUDY 2

This study was designed to investigate two practical questions that are of some importance and interest to test developers:

1. What are the effects of examinee sample size and test length on the standard errors of ability estimation curves?

2. What effects do the statistical characteristics of an item pool have on the precision of standard errors of ability estimation curves?

Method

Variables

Tests of three lengths were considered: 10, 20, and 80 items. Since a test with 10 items is about as short a test as is used in practice, the 10-test item length was selected to be studied; and the 80-item test was selected because the length represents about as long a test as is used in practice.

Ability scores were simulated to be normally distributed (mean = 0, SD = 1). This assumption was made to conform with an assumption made in Urry's (1974) item parameter estimation method, which was used (with slight modifications) in this study.

Three examinee sample sizes were used: 50, 200, and 1,000. The smallest sample size (N = 50) is considerably smaller than should be used in practice. It was chosen to identify the worst possible results that could be expected. The other two sample sizes define minimum and maximum sample sizes typically used in test development work with latent trait models. Ranges of parameter values for items in the two pools are shown in Table 6. As Table 6 shows, items in Item Pool 1 had a wider range of difficulty and discrimination values than those in Item Pool 2.

Table 6
Range of Item Parameter Values for the
Two Simulated Item Pools

Item Parameter	Range of Values	
	Item Pool 1	Item Pool 2
Difficulty (b)	-2.00 to 2.00	-1.00 to 1.00
Discrimination (a)	.60 to 2.00	.60 to 1.50
Pseudo-Chance (c)	.25 to .25	.25 to .25

Data Simulation

The eight steps in the data simulation were as follows:

1. Item Pool 1 was selected for study.

2. A test length (10, 20, or 80 items) and a sample size (50, 200, or 1,000 examinees) were selected. A sample of examinee ability scores were drawn from a normal distribution (mean = 0, SD = 1).

3. Computer program DATAGEN (Hambleton & Rovinelli, 1973), produced (1) item parameters, given the constraints of the item pool under investigation, and (2) examinee item scores. The computer program used the 3-parameter logistic model, the ability scores from Step 2, and item parameters generated at this step to produce probabilities of correct answers for examinees to the test items. These probabilities, in turn, were converted to examinee item scores (0 or 1) by a random number generator.

4. The examinee item scores from Step 3 were used in Urry's computer program to estimate item and ability parameters. However, only the item parameter estimates were used further in this particular study.

5. The item parameter estimates were used to obtain the standard errors of estimate for estimating θ [SEE (θ)]. The values of SEE(θ) at seven ability levels (θ = -3.00, -2.00, -1.00, 0.00, 1.00, 2.00, 3.00) were calculated.

6. Steps 3 to 5 were repeated three times to obtain three estimates of SEE(θ). All item and ability parameter values for the three runs were

identical. The particular examinee item scores varied from one run to the next because of the probabilistic nature of the score outcomes.

7. Steps 3 to 6 were repeated for each combination of test length and sample size (3 × 3 = 9).

8. Steps 2 to 7 were repeated with Item Pool 2. In all, 54 sets of test data were considered in the study.

Results

Tables 7 to 9 contain the SEE curves from Item Pool 1 obtained for three replications of three examinee sample sizes (N = 50, 200, and 1,000) and three test lengths (n = 10, 20, and 80) and for seven ability levels. Test lengths and sample sizes given under the column headed "Actual" are the number of items and examinees remaining after a satisfactory set of item and ability parameter estimates were obtained from Urry's computer program.

Effect of Sample Size

The data for a test length of 10 items, shown in Table 7, clearly show the lack of stability of the SEE curves for all sample sizes. There was little improvement, if any, due to increasing sample size. This result, however, may be due to the limited amount of data considered, since improvements were obtained in Item Pool 2 and at other test lengths.

Table 7
Standard Error Estimates (SEE) Adjusted to Correspond to 10-Item Tests for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool

Sample Size and Replication	Actual		Ability Level						
	Test Length	Sample Size	-3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50									
1	10	34	.66	.33	.67	.22	.75	1.60	2.19
2	10	34	2.40	1.88	.56	1.04	.20	1.34	1.37
3	9	34	.73	.57	1.03	.22	.58	.43	2.19
200									
1	10	172	.64	.21	.52	2.15	1.60	1.50	1.48
2	10	137	.22	.51	.36	1.30	.37	.96	2.45
3	10	174	2.63	2.14	.27	2.75	.92	.76	1.91
1000									
1	10	841	.98	.26	.58	1.43	3.33	.57	1.18
2	10	833	1.03	1.03	.67	1.05	.45	1.01	1.06
3	10	892	2.44	.49	.67	.30	.29	.89	1.33

Table 8 contains the results for 20 item test lengths and shows that the SEE curves were beginning to stabilize. Except at extreme values of the ability continuum, the results for the smaller sample sizes were nearly as good as those obtained with the larger sample size (N = 1,000).

Table 8
Standard Error Estimates (SEE) for Various Sample Sizes and Ability Levels with a Heterogeneous Item Pool

Sample Size and Replication	Actual		Ability Level						
	Test Length	Sample Size	-3.0	-2.0	1.0	0.0	1.0	2.0	3.0
50									
1	20	50	2.84	.70	.35	.30	.31	.44	1.23
2	20	50	1.93	1.53	.39	.32	.24	.45	1.19
3	20	46	2.07	.83	.58	.31	.36	.68	1.48
200									
1	20	193	—	.57	.26	.39	.33	.50	.77
2	20	196	—	1.51	.37	.34	.25	.53	.86
3	20	196	—	1.03	.22	.49	.34	.40	1.15
1000									
1	20	955	—	1.05	.48	.33	.33	.45	.82
2	20	969	—	1.18	.37	.33	.37	.40	.99
3	20	968	—	1.56	.40	.42	.32	.43	1.07

At a test length of 80 items, the SEE curves were highly stable, as clearly shown in Table 9. Similar to the effect noted with test lengths of 20, the expected decrease in variation of the SEE with increase in sample size was apparent only at ability levels of -1, +1, and +2.

Effect of Test Length

Examination of the results reported in Tables 7 through 9 indicate that for samples of size 50, as test length increased, variation in the SEE curves decreased at all ability levels. Results of the simulations for sample sizes of 200 and 1,000 clearly show the following trends:

1. The most stable SEE curves were obtained for the longest test length; and

2. For all ability levels, variation in the SEE curves decreased as test length increased.

Summary

Figure 1 illustrates the effect of test length and sample size on the stability of the SEE curves at five ability levels for Item Pool 1. Each graph represents a plot of the values of the SEE curves obtained when sample size was held constant and test length was varied. It is clear, from examination of these graphs, that sample size has little effect on the stability of SEE curves of the 10-item tests. The effect of sample size on the stability of the SEEs

Table 9
Standard Error Estimates (SEE) Adjusted to Correspond
to 80-Item Tests for Various Sample Sizes
and Ability Levels with a Heterogeneous Item Pool

Sample Size and Replication	Actual Test Length	Actual Sample Size	Ability Level -3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
50									
1	74	50	1.10	.35	.14	.14	.24	.24	.45
2	79	50	1.06	.48	.25	.17	.13	.32	.49
3	77	50	.93	.20	.19	.15	.17	.29	.48
200									
1	80	200	.89	.26	.22	.24	.19	.25	.44
2	80	200	.62	.29	.25	.19	.21	.25	.46
3	80	200	1.06	.35	.21	.19	.20	.25	.48
1000									
1	80	999	1.00	.35	.23	.21	.21	.24	.40
2	80	1000	.98	.32	.23	.22	.21	.23	.43
3	80	1000	1.08	.34	.20	.21	.20	.24	.46

was most apparent for the 20-item tests. For the 80-item tests sample size showed the most pronounced effect when there was an increase from 50 examinees (Figure 1a) to 200 examinees (Figure 1c). An effect was also noticed when sample size was increased from 200 examinees (Figure 1a) to 1,000 examinees (Figure 1c); however, the improvements in precision were more modest in size.

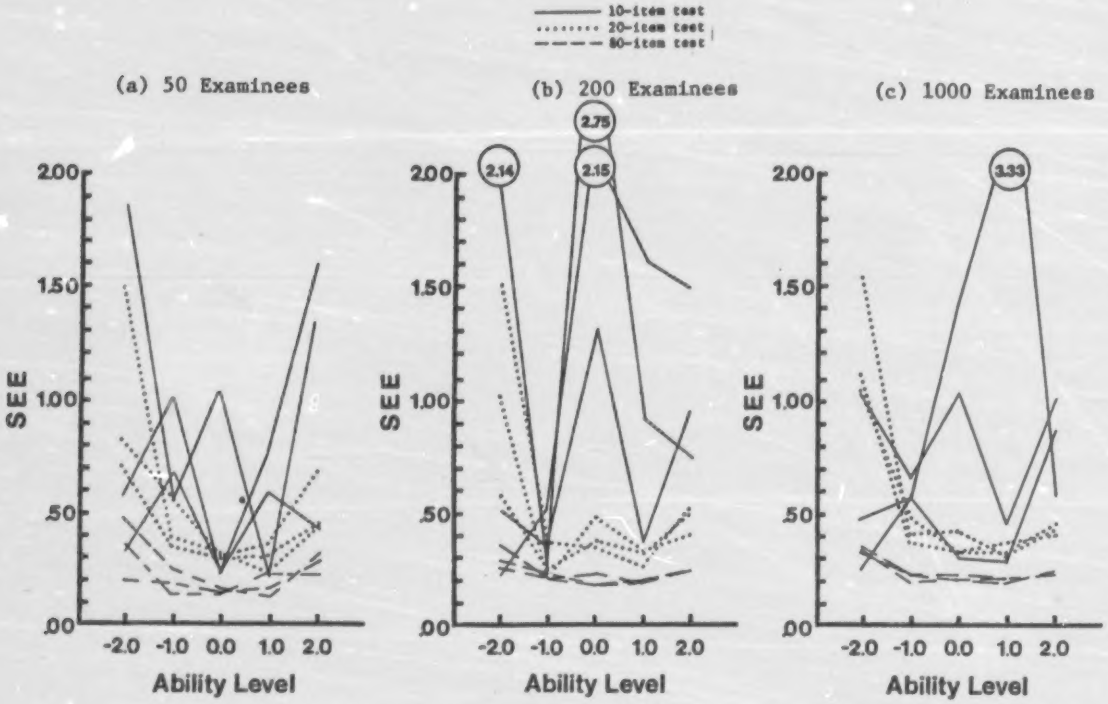
Table 10 summarizes the data reported in Tables 7 through 9 and includes summary data for Item Pool 2. Entries in this table are the standard deviations of the SEEs obtained across the three replications of the various studies. Standard deviations are reported for each test length-sample size combination across five ability levels. Also included in Table 10 is the average of the standard deviations across ability levels for each combination of test length and sample size.

Several trends are apparent from examination of the average variation of the SEEs for Item Pool 1: (1) the variation decreased as test length increased for all sample sizes; (2) when test length was fixed at 10 items, sample size had little or no effect on the stability of the SEE curves; and (3) sample size, generally, had a noticeable effect on the stability of the SEE curves.

Examination of the average variation across ability levels for Item Pool 2 indicated that for all test lengths, sample size has a noticeable effect on the stability of SEE curves. In comparison to the results reported for Item Pool 1, the effect of test length on the average variation across ability levels was not so apparent. The reason for this is the smaller variation observed for short tests with this particular item pool.

The results in Table 10 indicate that for tests of 20 and 80 items, the

Figure 1
Standard Errors of Estimate (SEE) for Three Test Lengths (10, 20, and 80 Test Items), Five Ability Levels and Three Sample Sizes (50, 200, and 1000 Examinees)
(Each Combination of Conditions Was Replicated Three Times)



(a) 50 Examinees (b) 200 Examinees (c) 1000 Examinees

variation in the SEE curves, averaged across ability levels, was very similar for both item pools. For test lengths of 10, the situation was quite different. In order to make the average variation across ability levels at this test length comparable for both item pools, these values were recomputed for Item Pool 2 excluding the values obtained for ability level of -2. The recomputed average variation values were .33, .38, and .52 for sample sizes of 50, 200, and 1,000, respectively. It is clear that for short tests, the homogeneous item pool (Item Pool 1) resulted in smaller average variations than did the heterogeneous item pool. A second point worth noting, is that the heterogeneous item pool (Item Pool 2) provided more stable SEEs at an ability of -2 for test lengths of 10 or 20 items than did the homogeneous item pool. For test lengths of 80, the results appear to be about the same for both item pools.

Conclusions

This study has provided data concerning the size of improvements in SEE curves relative to the three factors under investigation: (1) sample size, (2) test length, and (3) item pool characteristics. Several conclusions appear to be warranted:

Table 10
Standard Deviations of Standard Errors of Estimates Across Three Replications at Several Ability Levels for Different Test Lengths and Examinee Sample Sizes, and for the Heterogeneous Item Pool (Pool 1) and the Homogeneous Item Pool (Pool 2)

Test Length	Sample Size and Item Pool	Ability Level					Average Variation Across Ability Levels
		-2.0	-1.0	0.0	1.0	2.0	
10	50						
	Pool 1	.68	.20	.39	.23	.50	.40
	Pool 2		.17	.11	.41	.28	.24
	200						
	Pool 1	.85	.10	.60	.50	.31	.47
	Pool 2		.03	.07	.03	.22	.09
	1000						
	Pool 1	.32	.04	.47	1.40	.19	.60
	Pool 2		.07	.03	.04	.03	.04
20	50						
	Pool 1	.36	.10	.01	.05	.11	.16
	Pool 2	.78	.07	.10	.05	.08	.22
	200						
	Pool 1	.38	.06	.06	.04	.06	.12
	Pool 2	.37	.00	.02	.04	.00	.09
	1000						
	Pool 1	.22	.05	.04	.02	.02	.09
	Pool 2	.50	.03	.01	.00	.02	.11
80	50						
	Pool 1	.11	.04	.01	.05	.03	.06
	Pool 2	.16	.04	.01	.02	.04	.05
	200						
	Pool 1	.04	.02	.02	.01	.00	.02
	Pool 2	.03	.01	.01	.01	.01	.01
	1000						
	Pool 1	.01	.01	.00	.00	.00	.00
	Pool 2	.02	.00	.00	.00	.01	.01

1. Both test length and sample size are extremely important factors in the precision of SEE curves. The small number of reversals in the results was no doubt due to sampling fluctuations.

2. At the extremes of an ability continuum precision of SEE curves is very poor, even with large examinee sample sizes. The results are substantially better when tests are lengthened, even if the sample size is small (N = 50).

3. The precision of SEE curves would be acceptable in most instances if

the curves are based on 200 or more examinees with test lengths of at least 20 items. This recommendation holds if primary concern is with values of the curves in middle regions of the ability continuum [-1 to +1].

4. Increases in examinee sample sizes from 50 to 200 produce sizeable improvements in the precision of SEE curves; however, gains in precision due to increasing a sample size from 200 to 1,000 produce only modest gains in precision of the SEE curves.

5. Similarly for test lengths, improvements in precision were substantially better when the change was from 10 to 20 items than from 20 to 80 items.

The results of this study suggest that if an item pool is typical, the stability of SEE curves across readministrations of the test to similar groups of examinees will be quite good if the test includes at least 20 items and if 200 or more examinees are used in deriving the item statistics.

REFERENCES

Dinero, T. E., & Haertel, E. Applicability of the Rasch model with varying item discriminations. Applied Psychological Measurement, 1977, 1, 581-592.

Gulliksen, H. Theory of mental tests. New York: John Wiley & Sons, 1950.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R. K., & Rovinelli, R. A FORTRAN IV program for generating examinee response data from logistic test models. Behavioral Science, 1973, 18, 74.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hambleton, R. K., & Traub, R. E.. The robustness of the Rasch test model (Report No. 42). Amherst: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluative Research, 1976.

Hambleton, R. K., & Traub, R. E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Lord, F. M. Evaluation with artifical data of a procedure for estimating ability and item characteristic curve parameters (Research Bulletin 75-33). Princeton, NJ: Educational Testing Service, 1975.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Tinsley, H. E. A., & Dawis, R. Test-free person measurement with the Rasch simple logistic model. Applied Psychological Measurement, 1977, 1, 483-487.

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1968.

ACKNOWLEDGMENTS

This research was performed pursuant to a contract from the United States Air Force Office of Scientific Research. However, the opinions expressed here do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred. A complete report of Study 2 is contained in L. L. Cook & R. K. Hambleton, Effects of test length and sample size on the estimates of precision of latent ability scores (Report No. 87). Amherst: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluative Research, 1979. The authors are indebted to Janice Gifford for her extensive help in the collection and analysis of data reported in Study 2.

ESTIMATING ABILITIES WITHIN THE TWO-PARAMETER LOGISTIC LATENT TRAIT MODEL IN THE PRESENCE OF A NON-SYMMETRIC DISTRIBUTION OF ABILITY

MICHAEL WALLER
UNIVERSITY OF WISCONSIN--MILWAUKEE

Estimation of the parameters in the 2-parameter logistic latent trait model will be discussed within the framework of the estimation procedure developed by Bock and employed in the LOGOG computer program (Kolakowski & Bock, 1973). This method of estimation requires the assumption of some prior distribution of abilities during estimation of the item parameters (although no distributional assumption is required during estimation of the ability parameters). Typically, a normal prior is assumed during item parameter estimation. The questions to be explored in this monte carlo study are (1) What effect does this method of estimation have on the estimated abilities when the true distribution of abilities is nonsymmetric? and (2) Since the entire procedure is defined only to within a linear transformation, does there exist a linear function of the data that will improve the accuracy of the estimated abilities in this situation? The monte carlo simulation presented here reveals a plausible and simple candidate: a linear transformation using the means of the item difficulties and item discriminations. However, theoretical support in closed form for this solution is still forthcoming.

The motivation for seeking some function of the estimated item parameters to adjust the estimated abilities stems from the following well-known fact regarding the value of the discrimination parameter in the 1-parameter logistic model, commonly known as the Rasch model. The Rasch model is typically written as in Equation 1.

$$P_{ij} = \frac{e^{(\beta_j - \theta_i)}}{1 + e^{(\beta_j - \theta_i)}} \qquad [1]$$

Alternatively, it may be written as in Equations 2 and 3.

$$P_{ij} = \frac{e^{(\beta_j - \alpha_j\theta_i)}}{1 + e^{(\beta_j - \alpha_j\theta_i)}}; \ \alpha_j = \alpha \text{ for all } j; \qquad [2]$$

$$= \frac{e^{(\beta_j - \phi_i)}}{1 + e^{(\beta_j - \phi_i)}}; \; \phi_i = \alpha\theta_i \qquad [3]$$

Examination of Equations 2 and 3 makes it explicit that the 1-parameter Rasch model may in fact have item discrimination parameters that are all equal to some constant value, say α. The value of the constant will in most cases be unknown, as it will be considered in the variance of the distribution of the estimated abilities. Since this unknown item parameter may affect the distribution of the abilities, it is possible that unknown parameters of the distribution of abilities may affect the item parameters in a discernible way.

The Estimation Procedure

The entire estimation procedure is performed in two-step cycles. Estimation of abilities using the current item parameter estimates is the first step, and estimation of the item parameters using the current ability estimates is the second. In each cycle the mean and variance of the ability continuum are standardized to 0 and 1, respectively. The cycling continues until stable item parameters are reached.

Estimation of Ability

Estimation of abilities by maximum likelihood in this procedure, when specialized to binary choice data, is accomplished by the standard method as follows:

Let $j = 1, \ldots, n$ items;

$$r_{ij} = \begin{cases} 1 \text{ if person } i \text{ is correct on item } j; \\ 0 \text{ if person } i \text{ is incorrect on item } j; \end{cases}$$

θ_i = the latent ability of person i;

β_j = the item difficulty parameters; and

α_j = the item discrimination parameters.

Then, for a given person i the likelihood function is

$$L_i(\theta_i | r_i) = \prod_{j=1}^{n} P_{ij}^{\,r_{ij}} (1 - P_{ij})^{1 - r_{ij}}, \qquad [4]$$

where $P_{ij} = \Pr(r_{ij} = 1 | \theta_i)$; here

$$P_{ij} = \frac{e^{(\beta_j - \alpha_j\theta_i)}}{1 + e^{(\beta_j - \alpha_j\theta_i)}}. \qquad [5]$$

Therefore, the log likelihood is given by

$$\ell_i = \log\left[L_i(\theta_i | r_i\right] = \sum_j^n r_{ij} \log P_{ij}$$

$$+ (1 - r_{ij}) \log (1 - P_{ij}). \qquad [6]$$

Given the first and second derivatives of the log likelihood, Newton-Raphson iteration may be applied to the *k*th stage estimate of the parameter to yield the (*k* + 1)st stage estimate:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \frac{\partial \ell_i}{\partial \theta} \Bigg/ \frac{\partial^2 \ell_i}{\partial \theta^2} \qquad [7]$$

Estimation of Item Parameters

Estimation of the item parameters--difficulty, β_j, and discrimination, α_j--is not accomplished in the standard procedure (as described, for example, in Lord, 1968). Instead, the item parameters are estimated under the assumption that the abilities follow a previously specified distribution; here the normal distribution is used, with a mean and variance of 0 and 1, respectively. This is accomplished at each cycle by taking the current estimates of abilities, ranking them, and distributing them into 10 groups or fractiles in such a way that the number of subjects, N_i, across the *i* = 1,..., 10 fractiles reflect the normal distribution. Then, it is assumed that within each fractile the subjects are sufficiently homogeneous to permit proceeding as though there are N_i independent observations all at the same ability level, θ_i, some middle value in fractile *i*. Formally, the procedure is as follows:

Let *i* = 1, 2,..., *t* groups or fractiles whose subjects are sufficiently homogeneous as to be characterized by θ_i;
N_i = number of subjects in fractile *i* (determined by the assumed distribution);
r_{ij} = number of subjects in group *i* who respond to item *j* correctly;
r_j = $\{r_{ij}\}$ = vector of item responses to item *j* across the *t* fractiles;
β_j = the required difficulty parameter; and
α_j = the required slope parameter.

Then, for a given item *j* the likelihood function is

$$L_j(\beta_j, \alpha_j | \underline{r}_j = \prod_{i=1}^{t} \frac{N_i!}{r_{ij}!\,(N_i - r_{ij})!} P_{ij}^{\;r_{ij}} (1 - P_{ij})^{(N_i - r_{ij})} \qquad [8]$$

where

$$P_{ij} = Pr\,(r_{ij}\text{ correct }|N_i, \beta_j, \alpha_j, \theta_i) = \frac{e^{\beta_j = \alpha_j\theta_i}}{1 + e^{\beta_j - \alpha_j\theta_i}} \; .$$

Therefore,

$$\begin{aligned} \ell_j &= \log\left[L_j(\beta_j, \alpha_j | \underline{r}_j)\right] \\ &= C + \sum_i^t \left[r_{ij}\log P_{ij} + (N_i - r_{ij})\log(1 - P_{ij})\right] \qquad [9] \end{aligned}$$

Given the matrices of first and second derivatives, Newton-Raphson iteration may be applied to the k^th stage estimate of the parameters to yield the (k + 1)^st stage estimates:

$$\begin{bmatrix} \beta_j \\ \\ \alpha_j \end{bmatrix}_{k+1} = \begin{bmatrix} \beta_j \\ \\ \alpha_j \end{bmatrix}_k + \begin{bmatrix} \ell_{\beta_j\beta_j} & \ell_{\beta_j\alpha_j} \\ \\ \ell_{\alpha_j\beta_j} & \ell_{\alpha_j\alpha_j} \end{bmatrix}^{-1} \begin{bmatrix} \ell_{\beta_j} \\ \\ \ell_{\alpha_j} \end{bmatrix} \; . \qquad [10]$$

Example

The Data

The choice of the distribution of abilities for the monte carlo data was made to approximate an available set of data. The following distribution function was used:

$$F(\theta) = \frac{-\theta^2}{25} + \frac{\theta}{5} + \frac{3}{4} \; ; \qquad [11]$$

with corresponding density function

$$f(\theta) = \frac{-2\theta}{25} + \frac{1}{5} \; . \qquad [12]$$

The theoretical median, mean, variance, and coefficient of skewness were, respectively,

$$\text{Md} = \text{Median} = -1.03555 \qquad [13]$$

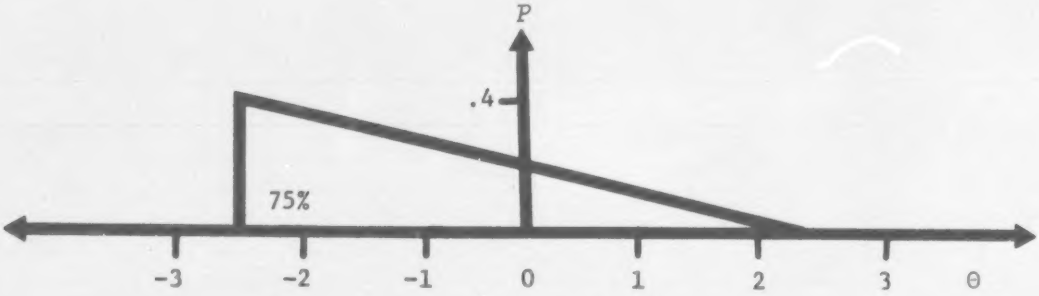
$$\mu = E(\theta) = -.8333\ldots \tag{14}$$

$$\sigma_\theta^2 = E[(\theta-\mu)^2] = 1.3888\ldots \tag{15}$$

$$\sqrt{\beta_1} = \frac{\mu_3^3}{\mu_2^{\frac{3}{2}}} = +4.4921. \tag{16}$$

and the density is approximated by Figure 1.

Figure 1
The Distribution of θ



The abilities ranged from -2.5 to +2.5, with 75% of the population lying between -2.5 and 0.

A sample of N = 480 abilities was generated by obtaining a random number in the unit interval for the value of the probability F and applying the inverse of the distribution function:

$$\theta = \frac{1 - \sqrt{1 + 4(3/4 - F)}}{2/5} . \tag{17}$$

For each simulated subject, responses to n = 45 items were generated. The difficulties of these items were set at values between -2.2 and +2.2 in steps of .1; the discriminations were all set equal to 1. Each subject's responses to these 45 items were generated by calculating the probability of a correct response, P_{ij}, using these item parameters and the subject's θ, and then comparing P_{ij} to a random number p in the unit interval. For each item

$$r_{ij} = \begin{cases} 1 & P_{ij} \geq p \\ 0 & P_{ij} < p \end{cases} \tag{18}$$

The criterion used to determine successful estimation of the sample's ability parameters was as follows: Construct approximate 95% confidence intervals around each subject's estimated ability using the estimate of the asymptotic variance of $\hat{\theta}_i$ given by the negative of the inverse of the second derivative of the log likelihood function. Then, simply count the number of subjects whose 95% confidence interval failed to cover the true ability and compare this number to the expected number from a binomial distribution with $\underline{p} = .05$.

Results

The results of the monte carlo study are as follows:

1. Estimating the abilities using the above procedure and placing a 95% confidence interval around each estimated ability yielded 353 out of the 480 simulated subjects for which the 95% confidence interval failed to cover the true ability.

2. The mean of the estimated item difficulties was $\bar{\underline{b}} = 0.898$; the mean of the estimated item discriminations was $\bar{\underline{a}} = 1.274$.

3. Applying the linear transformation

$$\hat{\theta}^*_i = \bar{b} + \bar{a}\,\hat{\theta}_i \qquad [19]$$

and the appropriate adjustments to the variance of the $\hat{\theta}$'s, yielding a standard error of $\sigma_{\theta^*_i} = \bar{\underline{a}}\sigma_\theta$ and then placing 95% confidence intervals around the transformed ability estimates, θ^*_i, yielded 31 out of 480 subjects for which the 95% confidence interval around the transformed ability estimate failed to cover the true ability--a result which did not differ significantly ($\underline{p} > .09$) from the expected number of 24 out of 480 subjects. In other words, with this transformation procedure, successful recovery of abilities was obtained.

Discussion

The results of this study should be neither over-interpreted nor under-interpreted. Although the study was based on only one sample of monte carlo data, the random number sequence utilized to generate these data was thoroughly checked for serial correlation and uniform distribution, utilizing the procedures presented in Hammersley and Handscomb (1964, chap. 3). Since the latent continuum was standardized to a mean of 0 and variance of 1 at each cycle, the shift in the mean of the difficulties as well as the shift in the mean of the discriminations cannot be interpreted simply as resulting from a failure to standardize the latent continuum.

Nevertheless, these are monte carlo results which are only at best loosely supported by theory. In addition, the behavior of this procedure in other circumstances in unknown, i.e., change the distribution of abilities or the distribution of either of the item parameters, and the adequacy of the procedure for recovering ability is undemonstrated. Consequently, extreme caution is recommended before utilizing the correction presented here.

The study does support the contention that there is an intimate connection between item parameter and ability parameter estimation. Although almost all estimation procedures in latent trait theory utilize the conditional two-step procedure--estimation of ability parameters followed by estimation of item parameters--estimation of the two sets of parameters is not independent. Consequently, latent trait methods that attempt to use a particular procedure in estimation but that begin by assuming, for example, that the item parameters are known and then present a "solution" to a particular problem for ability estimation, given known item parameters, are likely to be of limited practical utility.

REFERENCES

Hammersley, J. M., & Handscomb, D. C. Monte carlo methods. Norwich, Great Britain: Fletcher & Son, Ltd., 1964.

Kolakowski, D., & Bock, R. D. LOGOG: Maximum likelihood item analysis and test scoring--logistic model for multiple item responses. Chicago: National Educational Resources, 1973.

Lord, F. M. An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989 - 1020.

ESTIMATION OF PARAMETERS IN THE 3-PARAMETER LATENT TRAIT MODEL

HARIHARAN SWAMINATHAN AND JANICE GIFFORD
UNIVERSITY OF MASSACHUSETTS

The successful application of latent trait theory to practical measurement problems hinges upon the availability of procedures for the estimation of the parameters. Hence, investigations of the adequacy of the available procedures for estimating parameters in latent trait models are necessary and, indeed, play a crucial role when assessing the usefulness of latent trait theory.

Although the problem of estimating parameters in the 1-parameter latent trait model appears to be solved, some degree of controversy seems to surround the estimation of parameters in the 2- and 3-parameter models (Andersen, 1973; Wright, 1977). Lord (1975) has empirically evaluated the maximum likelihood procedure for estimating the parameters in the 3-parameter model and has provided answers to some of the questions that arise with respect to estimation of parameters. Jensema (1976) has compared the efficiency of a heuristic procedure suggested by Urry (1974) for estimating the parameters in the 3-parameter model with the maximum likelihood procedure. Ree (1979) has compared the properties of the Urry estimators and the maximum likelihood estimators and has investigated the effect of violating the underlying assumptions on the estimates, fixing the test length (80 items) and the number of examinees, however. Despite these efforts, little is known regarding the statistical properties of the estimators in the 3-parameter model and the effect of test length and examinee population size on the estimates.

Purpose

The purpose of this study was to investigate the efficiency of the Urry (1976) procedure and the maximum likelihood procedure for estimating parameters in the 3-parameter model, to study the properties of the estimators, and to provide some guidelines regarding the conditions under which they should be employed. In particular, the issues investigated were (1) the "accuracy" of the two estimation procedures, (2) the relationship between the number of items, examinees, and the accuracy of estimation, (3) the effect of the distribution of ability on the estimates of item and ability parameters, and (4) the statistical properties, such as bias and consistency, of the estimators.

Design of the Study

In order to investigate the issues mentioned above, artificial data were generated according to the 3-parameter logistic model

$$P_{ij}(\theta) = c_i + (1 - c_i)\ \{1 + \exp[-1.7\ a_i(\theta_j - b_i)]\} \qquad [1]$$

using the DATGEN program of Hambleton and Rovinelli (1973). Data were generated to simulate various testing situations by varying the test length, the number of examinees, and the ability distribution of the examinees. Test lengths were fixed at 10 items, 15 items, 20 items, and 80 items. Since the accuracy of maximum likelihood estimation with large numbers of items has been sufficiently documented by Lord (1975), tests with small numbers of items--10, 15, and 20--were chosen so that the accuracy of the estimation procedure could be ascertained for short tests. This is particularly important if latent trait theory is to be applied to criterion-referenced measurement. Similarly, the sizes of examinee population were set at 50, 200, and 1,000 in order to study the effect of small sample size on the accuracy of estimation.

In the Urry (1976) estimation procedure, the relationships that exist for item discrimination and item difficulty between the latent trait theory parameters and the classical item parameters are exploited (Lord & Novick, 1968, pp. 376-378). These relationships are derived under the assumption that ability is normally distributed and that the item characteristic curve (ICC) is the normal ogive. In order to study how the departures from the assumption of normally distributed abilities affect the Urry procedure, three ability distributions were considered: normal, uniform, and a negatively skewed distribution. The normal and uniform distributions were generated with mean 0.0 and variance of 1.0. (The uniform distribution was generated on the interval -1.73 to 1.73 to ensure unit variance.) A beta distribution with parameters 5 and 1.5 was generated to simulate a negatively skewed distribution, and then rescaled so that the mean was 0.0 and the variance 1.0. The distributions were standardized to remove the effect of scaling on the estimates of the parameters.

The three factors--test length (4 levels), examinee population size (3 levels), and ability distribution (3 levels)--were completely crossed to simulate 36 testing situations. Test data arising from these situations were subjected to the Urry estimation procedure using the computer program ANCILLES and to the maximum likelihood estimation procedure using the computer program LOGIST (Wood, Wingersky, & Lord, 1978).

Lord (1975) has emphasized the fact that simulated data should in some way resemble real data; otherwise, results obtained through simulation studies will not generalize to real situations. An attempt was therefore made to generate test data as realistically as possible. In order to accomplish this, item difficulty parameters were sampled from a uniform distribution defined in the interval $b = -2.0$ to 2.0, and item discrimination parameters were sampled from a uniform distribution in the interval $a = .6$ to 2.0. Since data were generated to simulate item responses to multiple-choice items with four choices, the pseudo-chance level parameters were set at $c = .25$. It should be noted, however, that this does not ensure close approximation of the generated data to real data. Combinations of item difficulty and discrimination that may not occur in constructed tests may occur with simulated tests and, hence, may affect the estimation procedures, limiting the generalizability of the findings in simulated studies to real situations. On the other hand, since the purpose of this study was to compare two estimation procedures and to study the statistical properties of estimators, the possible lack of correspondence between simulated and real data may not be a serious problem.

Results

Accuracy of Estimation

Comparisons between ANCILLES and LOGIST across various test lengths, examinee population sizes, and ability distributions are indicated in Tables 1, 2, and 3. The statistics reported are (1) the mean, μ, of the population item parameters for each population size; (2) the mean, $\overline{X}$, of the estimated item parameters; and (3) the correlation, ρ, between the true parameters and their estimates. These statistics are reported for the estimates obtained by employing both ANCILLES and LOGIST.

A comparison of the mean of the generated item parameters, μ, and the mean of the estimates, $\overline{X}$, for each of the item parameters--discrimination ($\underline{a}$), difficulty ($\underline{b}$), pseudo-chance level ($\underline{c}$), and the ability (θ) parameters--provides some indication of the accuracy of estimation. However, this comparison is rather weak when carried out alone, since the means do not contain all the essential information. Simultaneous comparisons of the means and examination of the correlations between the parameters and estimates, on the other hand, provide more complete information regarding the accuracy of estimation. If the correlation is high, and the means differ, then it can be concluded that the estimation was not sufficiently accurate.

Lord (1975) has implied that if heteroscedasticity exists, it may not be meaningful to compute correlations between true and estimated values, and, in general, the authors of this paper agree. However, since in the strict sense heteroscedasticity will invalidate the computation of a least squares regression line--the more appropriate criterion to employ is the generalized least squares criterion--and hence will rule out the use of simple, interpretable statistics for the evaluation of the accuracy of estimation, heteroscedasticity (when it occurred) was ignored; and correlations and least squares regression equations were computed.

Estimation of the discrimination parameter. Examination of the results in Tables 1, 2, and 3 indicates that the $\underline{a}$ parameter was poorly estimated for short tests. The highest correlation between true values and estimates for a test with 10 items and normally distributed ability was .36, with the mean of the estimates exceeding the mean of the true values. The correlations improved with increasing sample size and test length, with the mean of the estimated values approaching the mean of the true values from above. The highest correlation between the estimated and true values was .88 for an 80-item test with 1,000 examinees. This trend was also evident for the uniform and negatively skewed distributions of ability. In general, the $\underline{a}$ parameter was poorly estimated by ANCILLES, with the estimation improving more rapidly with increasing test length than with increasing examinee population size.

Table 1

Comparison of Estimates of Item and Ability Parameters from LOGIST and ANCILLES Based on a Normal Distribution of Ability

No. of Items	No. of Examinees	Discrimination (a)					Difficulty (b)					Chance-Level Parameter (c)					Ability (θ)				
			ANCILLES		LOGIST			ANCILLES		LOGIST			ANCILLES		LOGIST			ANCILLES		LOGIST	
		μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	SD	$\bar{X}$	SD	μ	$\bar{X}$	ρ	$\bar{X}$	ρ
10	50	1.46	3.47	.21	1.53	.43	-.15	-.87	.92	-.60	.95	.25	.34	.38	.12	.04	.02	-.10	.63	.00	.71
	200	1.18	2.82	.08	1.72	.46	.46	.41	.87	.22	.99	.25	.36	.18	.25	.00	.13	.07	.77	-.13	.76
	1000	1.46	2.97	.36	2.00	***	-.15	-.45	.95	-.15	.99	.25	.36	.28	.23	.02	-.00	.11	.71	.09	.75
15	50	1.17	2.08	-.25	1.67	-.02	.32	.61	.92	.29	.89	.25	.36	.25	.23	.00	.01	.04	.83	-.23	.78
	200	1.17	2.12	.38	1.59	.47	.32	.35	.97	.19	.96	.25	.35	.14	.23	.03	.11	-.00	.77	-.10	.77
	1000	1.40	2.61	.42	1.72	.86	-.09	-.04	.97	-.05	1.00	.25	.33	.17	.25	.00	.02	-.01	.86	-.04	.85
20	50	1.35	2.09	.40	1.60	.37	.16	.22	.95	.04	.96	.25	.29	.14	.18	.04	-.08	-.04	.87	-.00	.87
	200	1.35	2.17	.33	1.41	.46	.16	.22	.97	.08	.97	.25	.30	.12	.24	.01	-.02	.05	.88	-.12	.88
	1000	1.35	1.99	.66	1.59	.76	.16	.37	.98	.16	.99	.25	.36	.11	.25	.02	.00	.05	.89	-.06	.88
80	50	1.28	1.48	.40	1.40	.62	.15	.06	.85	.13	.88	.25	.20	.13	.22	.02	-.08	.12	.96	-.00	.97
	200	1.28	1.42	.64	1.46	.81	.15	.20	.96	.15	.98	.25	.22	.09	.25	.01	-.04	.09	.98	-.00	.97
	1000	1.28	1.36	.84	1.37	.88	.15	.21	.99	.12	1.00	.25	.23	.08	.25	.01	-.00	.08	.98	-.02	.97

***Indicates that correlation coefficient was not computed, since estimates of discrimination parameters attained the upper bound, 2.00.

Table 2

Comparison of Estimates of Item and Ability Parameters from LOGIST and ANCILLES Based on a Negatively Skewed Distribution of Ability

No. of Items	No. of Examinees	Discrimination (a)					Difficulty (b)					Chance-Level Parameter (c)					Ability (θ)				
			ANCILLES		LOGIST			ANCILLES		LOGIST			ANCILLES		LOGIST			ANCILLES		LOGIST	
		μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	SD	$\bar{X}$	SD	μ	$\bar{X}$	ρ	$\bar{X}$	ρ
10	50	1.18	2.67	.13	1.81	-.38	.46	.79	.91	.68*	.78*	.25	.39	.28	.25	.02	.05	.16	.70	-.30	.78
	200	1.46	2.52	-.58	1.96	-.31	-.15	.31	.98	-.19	.99	.25	.56	.26	.20	.01	-.07	-.01	.71	-.14	.78
	1000	1.46	2.98	-.06	1.95	-.31	-.15	.62	.97	-.26	.98	.25	.41	.40	.22	.02	-.01	.03	.57	-.17	.77
15	50	1.40	2.05	.10	1.11	-.01	-.09	.25	.96	-.31	.94	.25	.42	.16	.22	.01	-.17	-.13	.80	-.06	.82
	200	1.40	2.58	.23	1.61	.45	-.09	-.09	.94	-.37	.99	.25	.44	.27	.23	.01	.03	.08	.80	-.12	.91
	1000	1.40	2.37	.18	1.79	.87	-.09	.15	.95	-.10	1.00	.25	.48	.24	.25	.01	.00	.14	.79	-.06	.87
20	50	1.35	2.16	.49	1.25	.27	.16	.28	.92	.16*	.72*	.25	.34	.22	.19	.03	.02	.03	.81	-.14	.89
	200	1.35	2.03	.03	1.54	.10	.16	.21	.96	.05	.98	.25	.41	.19	.25	.00	.08	.15	.77	-.05	.87
	1000	1.35	2.10	.49	1.59	.52	.16	.51	.96	.08	.99	.25	.39	.13	.24	.01	.01	.06	.86	-.05	.91
80	50	1.29	1.49	.22	1.19	.28	.18	.10	.85	1.82*	.30*	.25	.21	.16	.21	.01	.07	.22	.93	-.30	.97
	200	1.28	1.26	.69	1.13	.61	.15	.22	.94	1.72*	.27*	.25	.21	.16	.24	.01	.14	.06	.96	-.04	.96
	1000	1.28	1.27	.68	1.24	.82	.15	.38	.97	.16	.99	.25	.20	.11	.25	.01	-.03	.08	.95	-.06	.97

*Indicates that the difficulty estimate for an item has taken on an extreme value.

Table 3
Comparison of Estimates of Item and Ability Parameters from LOGIST and ANCILLES Based on a Uniform Distribution of Ability

No. of Items	No. of Examinees	Discrimination (a)					Difficulty (b)					Chance-Level Parameter (c)					Ability (θ)				
			ANCILLES		LOGIST			ANCILLES		LOGIST			ANCILLES		LOGIST			ANCILLES		LOGIST	
		μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	ρ	$\bar{X}$	ρ	μ	$\bar{X}$	SD	$\bar{X}$	SD	μ	$\bar{X}$	ρ	$\bar{X}$	ρ
10	50	1.18	2.50	.33	1.26	.02	.46	.64	.68	.40	.81	.25	.43	.20	.18	.04	.06	.06	.59	.00	.71
	200	1.46	2.86	.60	1.74	.70	-.15	-.28	.90	-.49	.94	.25	.36	.19	.21	.00	.15	.09	.66	-.02	.75
	1000	1.46	2.52	.22	2.00	***	-.15	.06	.98	-.13	.99	.25	.33	.14	.29	.02	-.02	-.04	.74	-.10	.77
15	50	1.40	2.85	.33	1.90	.47	-.09	-.13	.91	-.04	.96	.25	.22	.16	.25	.01	.07	-.04	.90	-.00	.90
	200	1.40	2.70	.13	1.52	.03	-.09	-.04	.92	-.03	.91	.25	.22	.12	.20	.02	-.04	-.04	.89	-.00	.88
	1000	1.40	2.43	.37	1.61	.11	-.09	.16	.95	.20	.87	.25	.25	.09	.23	.01	-.03	-.02	.88	-.03	.87
20	50	1.35	2.35	.09	1.69	.47	.16	.52	.94	.24	.91	.25	.26	.30	.25	.01	-.07	.02	.89	-.14	.88
	200	1.35	2.08	.46	1.68	.34	.16	.40	.92	.07	.98	.25	.35	.24	.25	.00	.09	.05	.91	-.10	.88
	1000	1.35	1.98	.43	1.64	.56	.16	.34	.99	.06	1.00	.25	.46	.34	.24	.02	.04	.03	.90	-.02	.89
80	50	1.29	1.38	.30	1.38	.29	.18	.51	.88	.53	.86	.25	.20	.14	.21	.03	-.30	.09	.95	.00	.96
	200	1.28	1.32	.54	1.38	.73	.15	.29	.93	.20	.95	.25	.30	.15	.23	.01	-.04	.08	.97	.00	.97
	1000	1.28	1.26	.83	1.34	.94	.15	.22	.98	.12	1.00	.25	.36	.18	.25	.00	.02	.08	.97	-.00	.97

***Indicates that correlation coefficient was not computed, since estimates of discrimination parameters attained the upper bound, 2.00.

Table 4
Regression Coefficients and Standard Errors for Predicting the Estimates from True Values Based on a Normal Distribution of Ability

No. of Items	No. of Examinees	Discrimination (a)								Difficulty (b)								Ability (θ)							
		ANCILLES				LOGIST				ANCILLES				LOGIST				ANCILLES				LOGIST			
		b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE	b_0	SE	b_1	SE
10	50	2.55	1.38	.63	.95	.19	.90	.92	.62	-.71	.17	1.06	.15	.78	.13	1.20	.12	-.11	.17	.48	.11	-.02	.09	.77	.11
	200	2.57	.91	.21	.77	1.17	.36	.45	.27	-.04	.20	.97	.18	-.20	.06	.91	.04	-.01	.06	.59	.04	-.25	.04	.97	.06
	1000	1.52	1.22	.99	.83	***				-.29	.14	1.06	.11	.00	.07	1.03	.04	.11	.03	.58	.02	-.09	.02	.94	.03
15	50	2.89	.80	-.69	.68	1.72	.42	-.04	.35	.30	.14	.98	.11	-.01	.17	.94	.13	.03	.09	.78	.08	-.24	.07	1.00	.12
	200	1.23	.55	.76	.47	.96	.32	.54	.26	.09	.11	.80	.06	-.13	.11	1.01	.07	-.07	.05	.68	.04	-.20	.04	.91	.06
	1000	.85	.99	1.26	.71	.24	.23	1.06	.16	.05	.08	.95	.07	.05	.02	1.08	.00	-.03	.02	.78	.02	-.06	.01	.96	.02
20	50	1.14	.49	.70	.36	.83	.44	.57	.33	.07	.09	.95	.07	-.10	.08	.88	.06	-.10	.08	.82	.07	.07	.07	.87	.07
	200	.70	.99	1.09	.73	.49	.41	.68	.30	.04	.06	1.11	.06	-.09	.06	1.08	.07	.07	.03	.82	.03	.09	.03	1.07	.04
	1000	.36	.41	1.21	.30	.23	.27	1.01	.20	.21	.06	1.02	.05	.00	.04	1.02	.04	.05	.02	.80	.01	-.06	.01	.98	.02
80	50	.46	.27	.80	.21	.26	.16	.89	.13	-.06	.07	.83	.06	-.01	.06	.94	.06	.18	.07	.74	.03	.06	.05	.80	.03
	200	.11	.17	1.02	.14	.23	.10	.96	.08	.05	.04	.97	.03	.04	.03	.74	.02	.12	.02	.89	.01	.04	.02	.93	.02
	1000	.12	.09	.97	.07	.11	.08	.98	.06	.05	.02	1.03	.02	-.02	.01	.96	.00	.08	.01	.91	.01	-.02	.01	.96	.01

***Regression coefficients and their standard errors were not computed, since all estimates of discrimination parameters attained the upper bound, 2.00.

The least squares regression lines (for normally distributed ability) for predicting the estimates from true values, given in Table 4, were plotted (not shown) and compared with the line $y = x$ in order to determine the extent of the bias in estimation. The regression lines for all the test-length and sample-size combinations fell above the line $y = x$, indicating that ANCILLES systematically overestimated the a parameter, with the regression lines approaching the line $y = x$ with increasing test length. Again, the convergence to the line $y = x$ was more rapid with increasing test length than with increasing sample size.

Trends similar to that observed with ANCILLES were also observed with LOGIST. Although the estimation of a was poor, the LOGIST estimates were consistently better than those from ANCILLES in that the correlations between true values and estimates were higher and the means of the estimates were much closer to the means of the true values. Comparison of the plots of the regression lines, given in Table 4, with the line $y = x$ showed that although there was a general tendency for the parameters to be overestimated, this tendency was not as marked as with ANCILLES; the convergence of the regression lines to the line $y = x$ was more rapid. These trends--the higher correlations between true and estimated values than for ANCILLES estimates, the tendency for the means of the estimates to be closer to the means of the true values, and the rapidity of convergence of the regression line to the line $y = x$--were also observed with the uniform and negatively skewed distribution of ability.

Estimation of the difficulty parameter. ANCILLES was very successful in providing accurate estimates of the b parameter. The correlations between estimates and true values ranged from .85 to .99. Comparison of the regression lines for normally distributed ability, given in Table 4, with the line $y = x$ indicated that with the exception of tests with 10 items, the b parameter was generally overestimated for tests with 15 and 20 items. With larger numbers of items, there was a tendency for difficult items to be overestimated and for easy items to be underestimated. However, the bias was slight in that the convergence of the regression line to the line $y = x$ was rapid with increasing items and sample size.

In general, the LOGIST estimates of the b parameters were better than the estimates produced by ANCILLES. The correlations between true and estimated values ranged from .88 to 1.00, whereas ANCILLES yielded correlations ranging from .85 to .99. The means of the estimates were, in general, closer to the means of the true values than they were with ANCILLES. Comparisons of the regression lines, given in Table 4, with the line $y = x$ revealed that with increasing test length and sample size, the regression line approached the line $y = x$ rather rapidly, demonstrating that there was no bias in the estimation. No clear trends were visible with 10, 15, and 20 items, although the test with 10 items and 50 examinees produced overestimates of the b parameter. These results appeared to hold for both the uniform and negatively skewed distributions of ability, although with the skewed distribution there were two instances when the estimates of difficulty went out of bounds. These cases are indicated with an asterisk in Table 2. However, with 80 items and 1,000 examinees, the agreement between estimated values and true values was comparable to that obtained with normally distributed ability.

In general, the b parameter was estimated rather well by both LOGIST and ANCILLES. LOGIST fared surprisingly well with small numbers of items and examinees in comparison with ANCILLES, and in general produced better estimates (as determined by the correlations) than did ANCILLES.

Chance-level parameter. The true value of the chance-level parameter was set at $c=.25$ for all the items. Given this lack of variation among the true values, correlations between estimates and true values were not computed. Hence, only the mean of the true values, the mean of the estimates, and the standard deviation of the estimates are reported in Tables 1, 2, and 3.

ANCILLES clearly produced very poor estimates of the c parameter. The means of the estimates were consistently higher than the mean of the true values, with relatively large standard deviations. LOGIST estimates, on the other hand, were close to the true values, with small standard deviations. The mean LOGIST estimates ranged from .12 to .25 for normally distributed ability, from .19 to .25 for skewed distribution of ability, and from .18 to .25 for uniformly distributed ability. In comparison, ANCILLES yielded estimates that ranged from .20 to .36, .20 to .56, and .22 to .46, respectively, for the three distributions of ability.

Estimation of ability. An examination of Tables 1, 2, and 3 indicates a consistent pattern in the estimation of ability (θ) for both LOGIST and ANCILLES. The correlations between true values and estimates did not seem to be affected by increasing sample sizes for fixed test lengths. On the other hand, increasing the lengths of the test greatly affected the magnitude of the agreement between true values and estimates. This not surprising trend held for the three distributions of θ.

In general, it appears that although no differences existed between the ANCILLES and LOGIST estimates of θ for tests with 15 items or more, the LOGIST estimates fared better than the ANCILLES estimates for short tests with 10 items. This effect was more pronounced with the skewed ability distribution.

A closer examination of the two estimates by comparing the regression lines (obtained by regressing the estimates on the true values with the line $y = x$) indicated that, in general, ANCILLES underestimated θ for examinees with high true abilities and overestimated θ for examinees with low true abilities. This may partly be attributed to the fact that the c parameters were overestimated. No such trends were evident with the LOGIST estimates. These regression lines rapidly converged to the line $y = x$ with increasing test length.

Effect of Ability Distribution

A χ^2 test was used to determine if the uniform and the beta distributions deviated sufficiently from the normal. The beta distribution yielded a χ^2 value of 63.5 when the tails of the normal distribution were excluded and a value of 193.1 when the tails were included. The uniform distribution yielded a χ^2 value of 69.6 when tails were excluded and 307.7 when the tails were included. This indicates that both distributions deviated sufficiently from the normal, with the uniform distribution deviating even more than the beta distribution.

Comparisons of the results in Tables 1, 2, and 3 reveal that, in general, the beta distribution affected both estimation procedures, while the uniform distribution produced results similar to those obtained using a normal ability distribution. Although the beta distribution affected the estimation of $\underline{a}$ for both procedures and $\underline{c}$ and θ for ANCILLES, the estimation of $\underline{b}$ did not seem to be affected in either case. ANCILLES fared poorly with the skewed distribution in comparison to LOGIST in the estimation of the $\underline{a}$, $\underline{c}$, and θ parameters.

The estimates for the $\underline{a}$ parameter, resulting from both procedures, were negatively correlated with the true values for short tests. For longer tests, although estimates from both procedures improved, ANCILLES produced poor estimates in comparison to LOGIST. For an 80-item test with 1,000 examinees, a correlation of .68 was obtained using ANCILLES, as compared to a correlation of .82 obtained from LOGIST.

The estimates of the $\underline{c}$ parameters resulting from ANCILLES were extremely high for all tests except those of 80 items. The mean values ranged from .20 to .56 with the beta distribution, as compared to a range of .20 to .36 for the normal distribution of ability. The LOGIST estimates, on the other hand, were underestimated but comparable to those obtained using a normal distribution of ability.

The LOGIST estimates of ability resulting from using a skewed distribution of ability were as good as, and in some cases better than, the estimates obtained with a normal distribution. In contrast, ANCILLES with a skewed distribution resulted in poorer estimates. This effect held true even as sample size and test length increased.

Thus, ANCILLES estimates of θ, $\underline{a}$, and $\underline{c}$ parameters seemed to be affected more dramatically than the LOGIST estimates when ability had a skewed distribution. It should be noted that although the uniform distribution had a larger χ^2 value than the beta distribution, the results obtained with the uniform distribution of ability were similar to those obtained with the normal distribution. It is, then, not departures from normality but departures from symmetry and the unavailability of examinees in the lower tail of the ability distribution that affected the estimation procedure.

Statistical Properties of Estimation

Bias. If $\underline{g}$ is an estimator of γ, then $\underline{g}$ is an unbiased estimator of γ if

$$E(\hat{\gamma}) = \gamma, \qquad [2]$$

where E(·) is the expectation operator. This is a desirable property of estimators.

Schmidt (1977) has pointed out that the Urry procedure, developed by Urry in 1974, systematically overestimated the $\underline{a}$ parameter and underestimated the $\underline{b}$ parameter. Urry (1976) suggested a correction for this and incorporated this into the ANCILLES program, employed to estimate parameters in this study. Since it appears that for large numbers of items and examinees the estimates are unbiased (Lord, 1975), in order to study the effect of this correction on the estimates and to examine if the LOGIST estimates were unbiased a relatively short test of 20 items with 200 examinees was selected, response data were generated, and item parameters were estimated; this was replicated 20 times. Since the replications were obtained by generating sets of random examinees, the bias in the estimator of ability was not investigated.

The results of the replications are presented in Table 5, in which the true value, μ, of the 20 item parameters is given together with the mean estimate, $\bar{X}$, of the item parameters over 20 replications. The standard error and the t value obtained as

$$t = (\bar{X} - \mu)/SE \qquad [3]$$

are also given to indicate the degree of departure of the mean estimate from the true value.

ANCILLES clearly overestimated the a parameter, as did LOGIST. However, the bias in the LOGIST estimates did not appear to be as severe as the bias in the ANCILLES estimates. This finding is borne out in Figure 1, where the regression line for predicting $\bar{X}$ from μ is plotted for both ANCILLES and LOGIST and compared with the line $y = x$. The LOGIST regression line is closer to the line $y = x$ and shows that small values of a were overestimated, while very large values tended to be estimated accurately, partly due to the fact that an upper limit was imposed on the estimates. On the other hand, ANCILLES tended to overestimate large values, even more than small values, of a.

With item difficulty, LOGIST tended to underestimate easy items, while producing relatively accurate estimates of very difficult items (Figure 2). ANCILLES, on the other hand, tended to overestimate items with high b levels and to underestimate items with negative b levels. In general, ANCILLES seemed to produce biased estimates of b throughout the entire range.

Consistency. If g_n is an estimator of γ, g_n is a consistent estimator of γ if for any positive ε and η there is some N such that

$$\text{Prob}\,\{|g_n - \gamma| < \varepsilon\} > 1 - \eta,\ n > N. \qquad [4]$$

Consistency is a desirable property in that it ensures that an estimator tends to a definite quantity, which is the true value to be estimated.

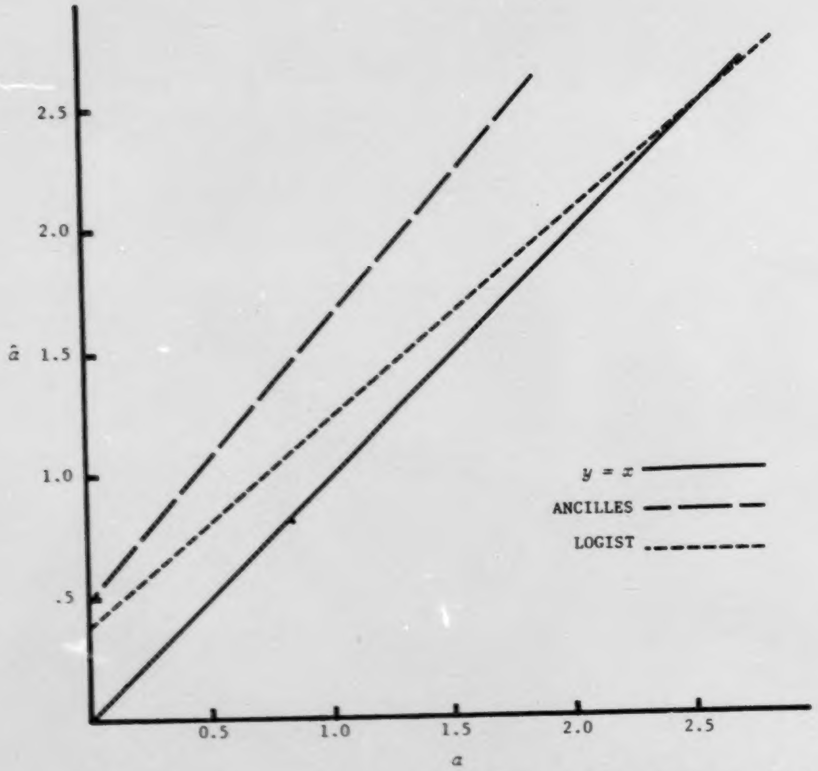
The problem of consistency has raised several questions concerning the estimation of parameters in the latent trait models. Andersen (1972) has argued that a consistent estimator of the discrimination parameter does not exist and, hence, has questioned the meaningfulness of the 2- and 3-parameter models.

In order to investigate whether or not the LOGIST and ANCILLES estimators were consistent, the regression equation for predicting the estimates from the true values of the various parameters were examined. The definition for a consistent estimator given earlier implies that an estimator is consistent if it is asymptotically unbiased and its variance tends to 0.0 with increasing sample

Table 5
Bias in the Estimation of Item Parameters
Based on 20 Replications for 20 Items and 200 Examinees

	Discrimination (a)							Difficulty (b)							Chance-Level Parameter (c)						
		ANCILLES			LOGIST				ANCILLES			LOGIST				ANCILLES			LOGIST		
Item	μ	X̄	SE	t	X̄	SE	t	μ	X̄	SE	t	X̄	SE	t	μ	X̄	SE	t	X̄	SE	t
1	.77	1.17	.067	5.9	.85	.073	1.9	1.63	1.90	.119	2.3	1.65	.097	.2	.25	.37	.020	6.0	.24	.007	-1.4
2	.84	1.60	.069	10.9	1.09	.129	1.9	-1.49	-1.28	.079	2.6	-1.52	.107	-.3	.25	.39	.017	8.2	.23	.004	-5.0
3	1.79	2.24	.056	8.04	1.89	.067	1.5	1.82	2.05	.112	2.1	1.75	.005	-1.3	.25	.33	.039	2.1	.23	.006	-3.3
4	1.11	1.91	.054	14.8	1.19	.119	.7	-1.54	-1.57	.076	-.4	-1.91	.188	-2.0	.25	.44	.021	9.0	.23	.004	-5.0
5	1.28	2.35	.202	5.3	1.57	.077	3.8	-.47	-.38	.206	.4	-.50	.057	-.5	.25	.33	.036	2.2	.23	.004	-5.0
6	1.53	2.42	.151	5.9	1.77	.088	2.6	-1.26	-1.24	.046	.4	-1.30	.063	.6	.25	.35	.036	2.8	.23	.004	-5.0
7	1.31	1.72	.129	3.2	1.62	.105	2.9	1.17	1.33	.079	2.0	1.06	.049	-2.2	.25	.30	.018	2.8	.23	.004	-5.0
8	1.31	1.89	.184	3.2	1.68	.098	3.8	1.47	1.75	.139	2.0	1.37	.067	-1.5	.25	.33	.036	2.2	.23	.005	-4.0
9	1.45	2.48	.117	8.8	1.54	.086	1.0	-1.78	-1.89	.089	-1.2	-1.96	.067	-2.7	.25	.48	.049	4.7	.23	.004	-5.0
10	1.48	2.23	.173	4.3	1.69	.088	2.4	-1.02	-.94	.053	1.5	-1.03	.044	-.2	.25	.34	.028	3.2	.23	.004	-5.0
11	1.58	2.26	.198	3.4	1.78	.081	2.5	.71	.83	.054	2.2	.67	.026	-1.5	.25	.29	.024	1.7	.23	.007	-2.8
12	1.43	1.89	.163	2.8	1.70	.106	2.6	.84	.97	.076	1.7	.74	.034	-2.9	.25	.30	.023	2.2	.23	.006	-3.3
13	1.97	3.06	.192	5.7	1.96	.027	-.4	.19	.07	.062	-1.9	.17	.023	-.9	.25	.18	.030	-2.3	.22	.007	-4.3
14	1.52	2.54	.133	7.7	1.62	.104	.9	-1.64	-1.72	.629	-.1	-1.89	.059	-4.2	.25	.41	.042	3.8	.23	.004	-5.0
15	.73	1.24	.105	4.9	.93	.112	1.8	.01	.22	.059	3.6	.01	.043	0.0	.25	.34	.019	4.7	.23	.006	-3.3
16	1.49	2.07	.221	2.7	1.75	.089	2.9	1.07	1.21	.073	1.9	1.03	.045	-.9	.25	.28	.026	1.2	.23	.005	-4.0
17	1.15	2.05	.129	6.9	1.47	.108	2.9	-.25	-.04	.045	4.7	-.23	.029	.7	.25	.35	.024	4.2	.22	.006	-5.0
18	1.89	2.29	.237	1.7	1.77	.089	-1.4	1.53	1.86	.135	2.4	1.57	.068	.6	.25	.31	.035	1.7	.23	.007	-1.4
19	1.23	1.58	.096	3.6	1.41	.127	1.4	1.28	1.77	.125	3.9	1.31	.086	.3	.25	.36	.016	6.9	.23	.004	-5.0
20	1.20	1.71	.076	6.7	1.56	.094	3.8	.94	1.05	.589	.2	.81	.068	-1.9	.25	.33	.010	8.0	.23	.007	-2.8

Figure 1
Bias in the Estimation of the Discrimination Parameter
of the 3-Parameter Logistic Model



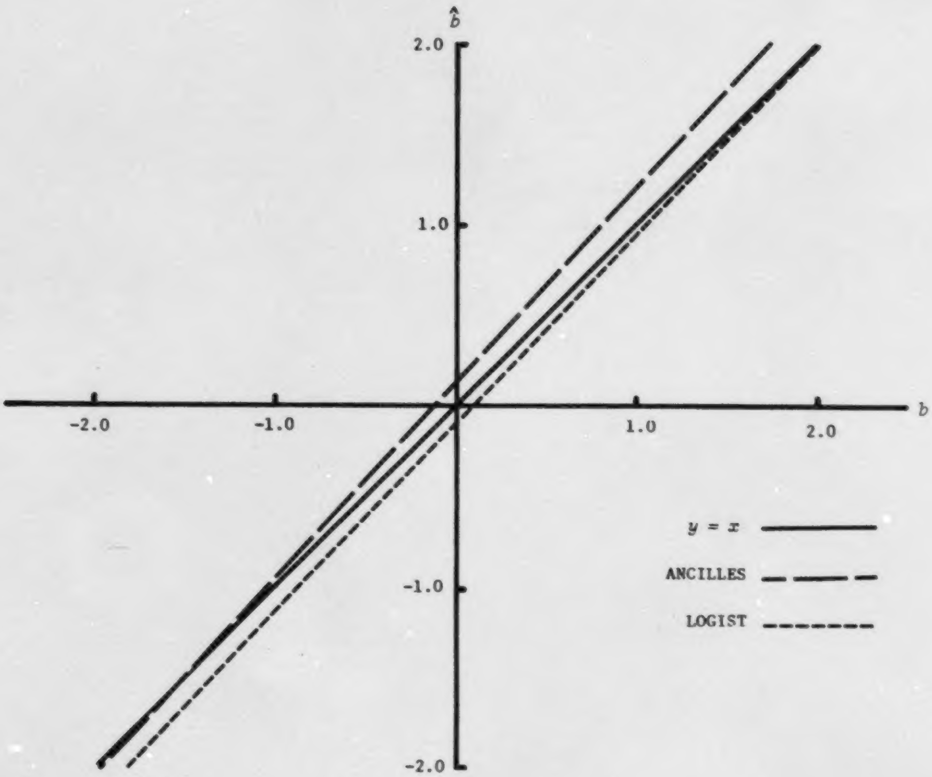
size. Consequently, in order for the estimators of the latent trait parameters to be consistent (1) the slope of the regression equation must approach 1.0 and the intercept must approach 0.0; and (2) the variance, and hence the standard errors of the estimate of the slope and intercept, must approach 0.0. If these conditions are met, then the estimator is consistent.

The regression coefficients and the standard errors are reported in Table 4. The results indicate that when both the number of items and the number of examinees increase, the slope and intercept coefficients approach 1.0 and 0.0, respectively, with the standard errors approaching 0.0. This tendency is evident for both ANCILLES and LOGIST estimators for the a, b, and c parameters, and for θ. In all these cases, the LOGIST estimator converged in probability to the true value more rapidly than the ANCILLES estimator. It should be pointed out, however, that the results reported here do not conclusively support this. It is clearly necessary to examine the standard errors and the regression coefficients with a greater number of items and examinees.

Discussion

The purpose of this study was to compare two methods for estimation of pa-

Figure 2
Bias in the Estimation of the Difficulty
Parameter of the 3-Parameter Logistic Model



rameters in the 3-parameter logistic model, the Urry method of estimation, and the maximum likelihood procedure. The computer programs that were used were the ANCILLES program and the LOGIST program (Wood, Wingersky, & Lord, 1978). The efficiency of the procedures were compared with respect to the accuracy of estimation, the effect of violating underlying assumptions (for ANCILLES), and the statistical properties of the estimators. The factors that were controlled were test length (4 levels), examinee population size (3 levels), and ability distribution (3 levels).

The results indicate that, in general, the maximum likelihood procedure was superior to the Urry procedure with respect to the estimation of all item and ability parameters. The differences were pronounced in the estimation of the discrimination and chance-level parameters, but with respect to the estimation of ability and difficulty parameters, the differences were less remarkable. Differing θ distributions had little effect on the estimation of $\underline{b}$ and θ. However, with a skewed distribution of θ, ANCILLES produced poorer estimates of $\underline{a}$ and $\underline{c}$ parameters than with normal or uniform θ distributions. LOGIST, although faring better than ANCILLES (with the exception of the 10-item test), produced slightly poorer results with the skewed distribution than with the normal or uniform distribution.

The number of examinees had a slight effect in improving the accuracy of estimation of the b and c parameters and θ. However, increasing the number of items and the number of examinees considerably improved the accuracy of the a estimates with both procedures. Surprisingly enough, a 20-item test with 1,000 examinees produced excellent estimates of the b and c parameters and reasonably good estimates of a and θ. Tests with 80 items and 1,000 people fared considerably better, providing good estimates of all parameters. Tests with 15 items or less, while yielding good estimates of b and c parameters and reasonable estimates of θ, yielded poor estimates of the a parameter. This severely limits the application of the 3-parameter latent trait model to criterion-referenced measurement situations, since criterion-referenced tests typically have fewer than 10 items. However, it should be pointed out that this limitation exists only if the item parameters and ability parameters are estimated simultaneously. If item banks with known item characteristics are employed to estimate ability, or if the 1-parameter model is employed, this limitation may not exist.

Although the LOGIST estimates were superior to the ANCILLES estimates, especially in the case of short tests, the difference between them was negligible when the number of items and the number of examinees increased. This is of particular importance, since ANCILLES requires considerably less computer time than LOGIST. The computer time taken by LOGIST, especially with large numbers of items and examinees, may become forbidding enough to warrant the use of ANCILLES in this situation. It should be noted that, in fairness to the maximum likelihood procedure, the Urry procedure, in general, deletes more items and examinees during estimation than does the maximum likelihood procedure. This may explain the rapidity of convergence and indicate a weakness in ANCILLES.

The bias and consistency results indicate that for small numbers of items, the estimates of the item and ability parameters are biased, with the ANCILLES more biased than the LOGIST estimates. As the number of examinees and the number of items increase, it appears that the estimators are unbiased and, in fact, are consistent. This, in a sense, supports a conjecture of Lord (1968) and shows that the 3-parameter model may be statistically viable.

REFERENCES

Andersen, E. B. Conditional inference in multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1973, *26*, 31-44.

Hambleton, R. K., & Rovinelli, R. A FORTRAN IV program for generating examinee response data from logistic test models. *Behavioral Science*, 1973, *18*, 74.

Jensema, C. A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement*, 1976, *36*, 705-715.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 1968, *28*, 989-1020.

Lord, F. M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters (Research Bulletin 75-33). Princeton, NJ: Educational Testing Service, 1975.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Ree, M. J. Estimating item characteristic curves. Applied Psychological Measurement, 1979, 3, 371-385.

Schmidt, F. L. The Urry method of approximating the item parameters of latent trait theory. Educational and Psychological Measurement, 1977, 37, 613-620.

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1977, 34, 253-269.

Urry, V. W. Ancillary estimators for the item parameters of mental tests. In W. A. Gorham (Chair), Computerized testing: Steps toward the inevitable conquest (PS-76-1). Washington, DC: U.S. Civil Service Commission, Personnel Research and Development Center, 1976. (NTIS No. PB 261 694)

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6). Princeton, NJ: Educational Testing Service, 1976. (Revised 1978)

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

ACKNOWLEDGMENTS

The project was performed pursuant to a contract from the United States Air Force Office of Scientific Research; however, the opinions expressed here do not necessarily reflect their position or policy, and no official endorsement by the Air Force should be inferred.

SMALL N JUSTIFIES RASCH METHODS

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

The usual Birnbaum item response function requires determining three parameters for each item; the Rasch model requires only one. If there is only a small group of examinees, the $\underline{a}$ parameter (the discriminating power) cannot be determined accurately for some of the items. The $\underline{c}$ parameters are even more of a problem. For small samples, is it perhaps better to use the Rasch model, estimating only one parameter per item, even though the Rasch model is incorrect?

For a better perception of the problem, consider a common prediction problem not related to item response theory: Suppose it is desired to predict variable $\underline{y}$ from measurements on five predictors. An available sample has been used to estimate the linear regression of $\underline{y}$ on the predictors. This regression equation may now be applied to estimate $\underline{y}$ for new individuals drawn from the same population. If the sample used to estimate the regression equation was large, the procedure is a good one; but if this sample was small, the procedure may be worse than simply using the sample mean of $\underline{y}$ as the predicted value of $\underline{y}$ for each new individual. Suppose, for example, that the true multiple correlation for predicting $\underline{y}$ was .40. If the sample had only 60 cases, the predictions from the sample regression equation would typically be no more accurate than a prediction that each new value of $\underline{y}$ will fall at the sample mean of $\underline{y}$.

It would be useful to know how large the sample of examinees must be before it is worthwhile to use a 2- or 3-parameter item response model in preference to the Rasch model. The answer to this question will, of course, depend on the purpose to be served. The present paper is a modest beginning: it only answers this question for the 2-parameter logistic model and only for one very limited situation. The purpose of this paper, then, is to point out the problem, to indicate a method of solution, and to provide some numerical results, indicating the sample size required when there is no guessing.

Method

Under the Rasch model, ability must be estimated by some function of the number-correct score $\underline{x}$, since this is a sufficient statistic under this model. Under the 2-parameter logistic model, ability must be estimated by a function of the weighted sum $\sum_{i=1}^{n} a_i u_i$ of item responses ($\underline{u_i}$), the weight for each item being the item discriminating power (a_i); under this model, this weighted sum is a sufficient statistic for estimating ability.

Given the a_i, the information function for number-correct score x and the information function for the weighted sum $\Sigma a_i u_i$ can be readily calculated and compared. The weighted sum always provides more information than the number-correct score except in the limiting case where the two scores are identical or proportional. In practice, the number-correct score perhaps provides up to 95% as much information as the weighted sum.

But now suppose that the a_i are not known but are only estimated. If the estimates $\hat{a}_i$ are sufficiently inaccurate, the weighted sum $\Sigma \hat{a}_i u_i$ will be less informative than the number-correct score x. The problem is to make a precise statement showing how the usefulness of the weighted sum $\Sigma \hat{a}_i u_i$ depends on the number of cases used to determine the estimated weights $\hat{a}_i$.

It is desired to compare $x \equiv \Sigma_i u_i$ and $\Sigma_i \hat{a}_i u_i$ as estimators of ability. Note, however, that expectations over the u_i for fixed $\hat{a}_i$ gives

$$\mathcal{E} x = \Sigma_i P_i(\theta) \quad , \qquad [1]$$

and

$$\mathcal{E} \Sigma_i \hat{a}_i u_i = \Sigma_i \hat{a}_i P_i(\theta) \quad , \qquad [2]$$

where $P_i(\theta)$ is the 2-parameter logistic item response function, the probability of answering the item correctly. This result shows that each scoring method provides an unbiased estimator of a function of ability, but the functions estimated are not the same.

A comparison must be made between a function of x and a function of $\Sigma_i \hat{a}_i u_i$ that estimate the same ability parameter. Moreover, the function of x should be independent of the a_i, since the estimation of a_i is not a part of any Rasch procedure. This will be done as follows:

The ability parameter to be estimated will be considered to be the individual's true number-correct score:

$$\xi \equiv \Sigma_i P_i(\theta) \quad . \qquad [3]$$

Note that this is simply a specified monotonic transformation of ability θ. Since x is an unbiased estimator of ξ, x is clearly the ideal statistic under the Rasch model for this purpose.

The optimal estimator of ξ under the 2-parameter model is not x. If there is no prior distribution for ξ, an optimal estimator is the function of the sufficient statistic $\Sigma_i a_i u_i$ that is unbiased for ξ. This function is uniquely determined by the Rao-Blackwell theorem (Kendall & Stuart, 1973, sec. 17.35), but it is too complicated for practical use here.

If the a_i and the item difficulties b_i are known, an asymptotically optimal estimator of ξ under the 2-parameter model is the maximum likelihood estimator (MLE),

$$\tilde{\xi} \equiv \Sigma_i P_i(\tilde{\theta}) \ , \qquad [4]$$

where $\tilde{\theta}$ is the MLE of θ when the a_i and b_i are known. This follows from the fact that the MLE of a given function of a parameter is, under regularity conditions, the same function of the MLE of the parameter. Moreover, this estimator $\Sigma_i P_i(\tilde{\theta})$ is actually a function of the weighted sum $\Sigma_i a_i u_i$, since the MLE is always a function of the sufficient statistic if such exists. In fact, $\tilde{\theta}$ is the solution of the likelihood equation

$$\Sigma_i a_i P_i(\tilde{\theta}) = \Sigma_i a_i u_i \qquad [5]$$

Since the a_i and b_i are not known, let $\hat{a}_i$ and $\hat{b}_i$, estimated from some previously available sample of examinees, be substituted. Thus, the 2-parameter estimator to be compared with x is

$$\hat{\xi} \equiv \sum_{i=1}^{n} \hat{P}_i(\hat{\theta}) \ , \qquad [6]$$

where $\hat{P}_i(\hat{\theta})$ is the item response function, with $\hat{a}_i$ and $\hat{b}_i$ substituted for the unknown true item parameters and $\hat{\theta}$ is the solution of

$$\Sigma_i \hat{P}_i(\hat{\theta}) - \Sigma_i \hat{a}_i u_i = 0 \ . \qquad [7]$$

If N is large enough, $\hat{\xi}$ will necessarily show the same advantage over x as does the weighted sum $\Sigma_i a_i u_i$ when the a_i are known. But what if N is small, so that the $\hat{a}_i$ and $\hat{b}_i$ are erroneous estimates?

Since x and $\hat{\xi}$ are both consistent estimators of the same ability parameter (ξ), they are properly compared by their mean squared errors (MSE). The exact sampling variance of x is

$$\mathrm{Var}(x) = \sum_{i=1}^{n} P_i(\theta) Q_i(\theta) \ . \qquad [8]$$

Since x is unbiased for ξ, this is also the exact MSE.

The sampling variance of $\hat{\xi}$ depends on ξ, the true score of the examinee whose true score is to be estimated. Given ξ, the variance of $\hat{\xi}$ arises from two sources:

1. Sampling fluctuations in the data on N examinees used to estimate the a_i and the b_i,
2. Sampling fluctuations in the responses (u_{ia}) of examinees at the given true-score level ξ.

The examinee whose true score is to be estimated is *not* included in the sample of N examinees; thus, the second source of error is independent of the first.

It does not seem feasible to obtain the exact MSE of $\hat{\xi}$ when $N < \infty$; consequently, the present study deals only with its asymptotic variance, which is equal to its asymptotic MSE. Formulas for calculating the asymptotic sampling variance are given in the Appendix.

Table 1
Item Serial Numbers and Item Parameters for All Tests Studied

Item Serial No.	Item Parameters a	b
3	1.6	-1.9
4	1.7	-1.5
5	0.8	-1.7
8	1.3	-1.7
9	0.4	0.5
10	1.1	-1.3
13	1.4	-1.2
14	0.9	-1.1
15	0.6	-1.9
18	1.2	-1.0
19	1.6	-0.9
20	0.6	-0.4
23	0.6	-1.3
24	0.5	-1.4
25	0.9	-0.9
28	1.8	-0.9
29	0.9	-0.8
30	0.5	0.3
33	0.7	-0.8
34	0.7	-0.4
35	1.0	-0.2
38	0.8	0.0
39	1.1	-0.4
40	0.8	0.1
43	0.5	0.8
44	1.1	-0.3
45	0.7	0.3
48	0.6	0.8
49	0.4	0.3
50	0.7	0.9

Test Studied

Numerical results can only be obtained for particular numerical values of the item parameters a_i and b_i. The following procedure was used in the hope of obtaining realistic numerical values.

The responses of 3,000 6th-grade students to a 50-item Metropolitan (MAT) vocabulary test were analyzed by LOGIST. Since (for simplicity) the present study was limited to the 2-parameter model, all c parameters were held at 0. The $\hat{a}_i$ and $\hat{b}_i$ obtained were used as true item parameters for the tests to be studied here. These item parameters are listed in Table 1.

Table 2 shows how 4 different 10-item tests are defined in terms of the items listed in Table 1. Tests 3, 4, and 5 are nonoverlapping spaced samples of items. Since the items in Table 1 are arranged roughly in order of difficulty, in Table 2 test difficulty tends to increase from top to bottom. Table 2 also shows for each test the true test score ξ that corresponds to specified values of θ. Remember that for any given test, ξ and θ are equivalent measures of the same ability, differing only in scale.

Table 2
True Score (ξ) Equivalent to Specified Ability Levels (θ)
for Four 10-Item Tests

		Specified Values of θ				
Test	Items in Test	-2	-1	0	1	2
3	3, 8, 13, 18,..., 48	1.8	4.8	7.4	8.7	9.4
4	4, 9, 14, 19,..., 49	1.5	4.1	7.1	8.7	9.3
5	5, 10, 15, 20,..., 50	1.7	3.8	6.3	8.2	9.3
1B	10, 10, 20, 20,..., 50, 50	1.2	3.1	5.4	7.4	8.8

Results

Number-correct score x is an unbiased estimator of ξ. On the other hand, $\hat{\xi}$ is only asymptotically unbiased. The exact small-sample bias of $\hat{\xi}$ was calculated for 10-item Test 1B and for 5-item Test 1A, parallel to Test 1B except for length. (The method used for computing $\mathcal{E}(\hat{\xi} - \xi|\xi)$ is entirely parallel to the method for computing $\text{Var}(\hat{\xi}|\xi)$ described in the Appendix.) Test 1B consisted of 2 items exactly like Item 10, 2 like Item 20, and so forth, for a total of 10 items. Test 1A consisted simply of Items 10, 20, 30, 40, and 50.

Table 3 compares the bias of these two tests that differed only in length. The bias was small, even for five-item tests. The true score ξ of Test 1B was exactly double the true score of Test 1A, but the bias in $\hat{\xi}$ increased more modestly, if at all, as the test length was doubled.

Table 4 shows the exact small-sample variance of $\hat{\xi}$ when the item parameters were known, determined from an infinitely large sample of examinees. In this

Table 3
Bias $(\mathcal{E}\hat{\xi} - \xi)$ in True Score Estimate $\hat{\xi}$ for Tests 1A and 1B, Which Were Parallel Except for Length, When Item Parameters Were Known $(N = \infty)$

	Test	
θ	1A (*n*=5)	1B (*n*=10)
-2	-.028	-.035
-1	.029	.037
0	.045	.047
1	.029	.026
2	.020	.020

table Tests 1A, 1B, and 1C, which were parallel except for length, are compared. Test 1C contained 3 items like Item 10, 3 like Item 20, and so forth, for a total of 15 items. As might be expected, the sampling variance increased almost exactly as test length increased.

Table 4
Variance (Equation A2) of True Score Estimate $\hat{\xi}$ When Item Parameters Were Known $(N = \infty)$ for Tests 1A, 1B, 1C, Which Were Parallel Except for Length

	Test		
θ	1A (*n*=5)	1B (*n*=10)	1C (*n*=15)
-2	.52	1.01	1.50
-1	.92	1.75	2.58
0	.93	1.87	2.82
1	.79	1.60	2.40
2	.47	.95	1.44

As noted previously, the optimal estimator of ξ is the function of $\Sigma_i a_i u_i$ that is unbiased for ξ. Since the desired function (which can be found by the Rao-Blackwell theorem) is impractical to use, the consistent estimator $\hat{\xi}$ was used. The MSE is equal to the variance plus the square of the bias. For Tests 1A and 1B, it can be seen from Tables 3 and 4 that the MSE differed from the variance of $\hat{\xi}$ only in the third decimal place. Table 5 compares the variance of x with the variance of $\hat{\xi}$. Replacing variance by MSE would not change the picture. The relative efficiency of two consistent estimators is asymptotically proportional to the ratio of their sampling variances. A comparison of the last 2 columns of Table 5 shows that the efficiency of x ranged from .85 at $\theta = 0$ and $\theta = -1$ to .93 at $\theta = -2$.

Table 5
Variance (Equation A2) of True Score Estimate $\hat{\xi}$ for Test 3, as a Function of the Sample Size (N) Used to Estimate the Item Parameters; Also Variance (Equation 8) of Number-Correct Score $\underline{x}$

θ	ξ	Var($\hat{\xi}\mid\xi$) when N=100	N=300	N=1,000	N= ∞	Var ($\underline{x}\mid\xi$)
-2	1.8	1.32	1.23	1.20	1.19	1.28
-1	4.8	1.84	1.76	1.73	1.72	1.90
0	7.4	1.18	1.13	1.12	1.11	1.30
1	8.7	.78	.74	.73	.72	.85
2	9.4	.47	.45	.44	.44	.50

Interpolating in Table 5, it can be seen that for $\theta = -2$, $\underline{x}$ was better than $\hat{\xi}$ when the item parameters were estimated from a sample with $N \overline{<} 200$, to a rough approximation; $\hat{\xi}$ was better than $\underline{x}$ when $N > 200$. It can be said, therefore, that N=200 is the *critical* sample size. For the other tabled θ values, the critical sample size is in each case less than 100.

The critical N's for Test 3 are listed in Table 6 along with similar values for Tests 4, 5, 1A, 1B, and 1C. Because of the heavy cost in computer time, no runs were made for 15-item tests other than Test 1C. It appears that for the 10- and 15-item tests, the Rasch estimator $\underline{x}$ may be slightly superior to the 2-parameter estimator $\hat{\xi}$ when the number of cases available for estimating the item parameters is less than 100 or 200. This is the main conclusion of the study.

Table 6
Approximate Number of Cases (N) Required for $\hat{\xi}$ To Have a Smaller Sampling Variance Than Number-Correct Score $\underline{x}$

θ	Test 1A	3	4	5	1B	1C
-2	700	200	<100	300	250	200
-1	3000	<100	<100	250	200	100
0	<100	<100	<100	200	<100	<100
1	100	<100	<100	150	150	200
2	100	<100	<100	100	250	250

Conclusions

This study has been limited to a comparison of 1-parameter (Rasch) and 2-parameter estimators of the examinee's true score. Similar studies should be made for the 3-parameter model. Estimators of other quantities, such as item difficulty, should also be compared. The same approach can, in principle, be applied to determine the relative effectiveness of the Rasch and other models for test equating and other practical purposes; however, the computational burden of doing this may prove to be excessive.

REFERENCES

Kendall, M. G., & Stuart, A. The advanced theory of statistics (Vol. 2; 3rd ed.). New York: Hafner, 1973.

APPENDIX

Asymptotic Sampling Variance of $\hat{\xi}$

By a standard formula from analysis of variance, the error variance of $\hat{\xi}$ for fixed ξ can be written

$$\mathrm{Var}(\hat{\xi}|\xi) = \mathcal{E}_{\underset{\sim}{u}}[\mathrm{Var}(\hat{\xi}|\xi,\underset{\sim}{u})|\xi] + \mathrm{Var}_{\underset{\sim}{u}}[\mathcal{E}(\hat{\xi}|\xi,\underset{\sim}{u})|\xi] \quad . \qquad [A1]$$

where $\mathcal{E}_{\underset{\sim}{u}}$ and $\mathrm{Var}_{\underset{\sim}{u}}$ are taken across all possible response vectors $\underset{\sim}{u} \equiv \{u_i\}$. To understand the terms in brackets, note that when $\underset{\sim}{u}$ is fixed, the only other source of variability is sampling error in the estimation of the a_i and the b_i. Remember that the $\hat{a}_i$ and $\hat{b}_i$ are obtained from a sample of N examinees and that $\underset{\sim}{u}$ belongs to an examinee who is not part of that sample.

For large N the last term in Equation A1 is adequately approximated by replacing the estimates $\hat{a}_i$ and $\hat{b}_i$ by their true values a_i and b_i; in other words, $\mathcal{E}(\hat{\xi}|\xi,\underset{\sim}{u})$ can be replaced by $\mathcal{E}(\tilde{\xi}|\xi,\underset{\sim}{u})$. By Equations 4 and 5, fixing $\underset{\sim}{u}$ also fixes $\tilde{\xi}$, so now $\mathcal{E}(\tilde{\xi}|\xi,\underset{\sim}{u}) \doteq \xi$. Thus, the last term in Equation A1 becomes $\mathrm{Var}_{\underset{\sim}{u}}(\tilde{\xi}|\xi)$. By Equation 3 whenever ξ is fixed, θ is fixed also; so Equation A1 can be written approximately

$$\mathrm{Var}(\hat{\xi}|\xi) = \mathcal{E}_{\underset{\sim}{u}}[\mathrm{Var}(\hat{\xi}|\theta,\underset{\sim}{u})|\theta] + \mathrm{Var}_{\underset{\sim}{u}}(\tilde{\xi}|\theta) \quad . \qquad [A2]$$

The first variance on the right arises from sampling fluctuations in the $\hat{a}_i$ and the $\hat{b}_i$; the second variance is independent of these fluctuations. The second variance can be evaluated as follows:

1. For each possible item response pattern $\underset{\sim}{u}$, determine $\tilde{\theta}$ by solving Equation 5 numerically.
2. For each $\tilde{\theta}$ from Step 1, compute $\tilde{\xi} \equiv \Sigma_i P_i(\tilde{\theta})$.
3. For each $\underset{\sim}{u}$, compute

$$\text{Prob}(\underset{\sim}{u}|\theta) \equiv \prod_{i=1}^{n} P_i(\theta)^{u_i}\, Q_i(\theta)^{1-u_i} \qquad [A3]$$

for the given values of θ (not $\tilde{\theta}$). This gives the frequency distribution of $\underset{\sim}{u}$.

4. Compute the variance for given θ of the $\tilde{\xi}$ obtained in Step 2, taken over the frequency distribution of $\underset{\sim}{u}$ found in Step 3. The result is $\text{Var}_{\underset{\sim}{u}}(\tilde{\xi}|\theta)$, as required.
5. Repeat the foregoing for different given values of θ, as desired.

Although the notation does not make it explicit, the results obtained depend, of course, on the $\underline{a}_i$ and $\underline{b}_i$ of the items in the test being studied. A separate study must be carried out for each different test. Because of the number of calculations required, practical considerations limit investigation to tests not much longer than 15 items.

It remains to evaluate the first term on the right of Equation A2. The quantity $\text{Var}(\hat{\xi}|\theta,\underset{\sim}{u})$ will be evaluated by the delta method. Let $\xi'_{\hat{a}i}$ denote the partial derivative of $\hat{\xi}$ with respect to $\hat{\underline{a}}_i$, and similarly for $\hat{\underline{b}}_i$ and $\hat{\theta}$. The total derivative of $\hat{\xi}$ is

$$d\hat{\xi} = \xi'_{\hat{\theta}}\, d\hat{\theta} + \Sigma_i \xi'_{\hat{a}i}\, d\hat{a}_i + \Sigma_i \xi'_{\hat{b}i}\, d\hat{b}_i \quad . \qquad [A4]$$

but now, however, by Equation 7, $\hat{\theta}$ is itself a function of the $\hat{\underline{a}}_i$ and the $\hat{\underline{b}}_i$. Denoting the left side of Equation 7 by $\hat{\ell}$, the total derivative of Equation 7 is

$$\ell'_{\hat{\theta}}\, d\hat{\theta} + \Sigma_i \ell'_{\hat{a}i}\, d\hat{a}_i + \Sigma_i \ell'_{\hat{b}i}\, d\hat{b}_i = 0 \quad , \qquad [A5]$$

where $\ell'_{\hat{a}i}$ is the partial derivative of $\hat{\ell}$ with respect to $\hat{\underline{a}}_i$, and similarly for $\hat{\underline{b}}_i$ and $\hat{\theta}$. Eliminating $\underline{d\hat{\theta}}$ from Equations A4 and A5 gives

$$\ell'_{\hat{\theta}}\, d\hat{\xi} = \Sigma_i(\ell'_{\hat{\theta}}\xi'_{\hat{a}i} - \xi'_{\hat{\theta}}\ell'_{\hat{a}i})\, d\hat{a}_i + \Sigma_i(\ell'_{\hat{\theta}}\xi'_{\hat{b}i} - \xi'_{\hat{\theta}}\ell'_{\hat{b}i})\, d\hat{b}i \qquad [A6]$$

When the delta method is applied to Equation 6, it is found that

$$\begin{aligned}\text{Var}(\hat{\xi}|\theta,\underset{\sim}{u}) = \frac{1}{\ell'^{2}_{\theta}}\, [\, &\Sigma_i\{(\ell'_{\theta}\xi'_{ai} - \xi'_{\theta}\ell'_{ai})^2\ \text{Var}(\hat{a}_i|\theta,u)\} \\ &+ \Sigma_i\{(\ell'_{\theta}\xi'_{bi} - \xi'_{\theta}\ell'_{bi})^2\ \text{Var}(\hat{b}_i|\theta,\underset{\sim}{u})\} \\ &+ 2\ \Sigma_i\ \{(\ell'_{\theta}\xi'_{ai} - \xi'_{\theta}\ell'_{ai})(\ell'_{\theta}\xi'_{bi} - \xi'_{\theta}\ell'_{bi}) \\ &\qquad \text{Cov}(\hat{a}_i,\hat{b}_i|\theta,\underset{\sim}{u})\}] \ . \end{aligned} \qquad [A7]$$

For given θ and $\underset{\sim}{u}$, the variance needed for the first term on the right of Equation A2 can be computed from Equation A7. The necessary derivatives are

$$\xi'_{\theta} = \Sigma_i a_i \pi'_i \quad , \tag{A8}$$

$$\xi'_{ai} = (\theta - b_i)\pi'_i \quad , \tag{A9}$$

$$\xi'_{bi} = -a_i \pi'_i \quad , \tag{A10}$$

$$\ell'_{\theta} = -\Sigma_i a_i^2 \pi'_i \quad . \tag{A11}$$

$$\ell'_{ai} = u_i - P_i - a_i(\theta - b_i)\pi'_i \quad , \tag{A12}$$

$$\ell'_{bi} = +a_i^2 \pi'_i \quad , \tag{A13}$$

where $\pi'_i = DP_iQ_i$ denotes the derivative of P_i with respect to $L_i \equiv \underline{a_i}(\theta - \underline{b_i})$, and $D = 1.7$. The variance-covariance matrix of $\underline{\hat{a}_i}$ and $\underline{\hat{b}_i}$, needed in Equation A7, is found by inverting the Fisher information matrix:

$$\begin{bmatrix} D^2 \sum\limits_{a=1}^{N} (\theta_a - b_i)^2 P_{ia} Q_{ia} & -D^2 a_i \sum\limits_{a=1}^{N} (\theta_a - b_i) P_{ia} Q_{ia} \\ -D^2 a_i \sum\limits_{a=1}^{N} (\theta_a - b_i) P_{ia} Q_{ia} & D^2 a_i^2 \sum\limits_{a=1}^{N} P_{ia} Q_{ia} \end{bmatrix} \tag{A14}$$

where $P_{ia} \equiv P_i(\theta_a)$.

To evaluate Equation A2 for fixed θ, compute $\mathrm{Var}(\xi|\theta,\underset{\sim}{u})$ by Equation A7 separately for each $\underset{\sim}{u}$. Then, take the average of these values across $\underset{\sim}{u}$ (weighting each value by $\mathrm{Prob}(\underset{\sim}{u}|\theta)$ given by Equation A3) to find the first term on the right. Add on the second term $\mathrm{Var}_{\underset{\sim}{u}}(\tilde{\xi}|\theta)$, computed as described earlier. The resulting $\mathrm{Var}(\hat{\xi}|\xi)$ must be computed separately for each ξ or θ of interest. (Note again that fixing ξ is equivalent to fixing θ, because of Equation 3.)

DISCUSSION: SESSION 8

BERT F. GREEN, JR.
JOHNS HOPKINS UNIVERSITY

The papers presented in this session seem to have been done competently and to have given reasonable results. I should like, however, to put their results in some perspective.

Why is latent trait theory attractive? It promises to deliver a scale that is essentially invariant over different item selections; therefore, the measurement scale provided, the θ scale, is paramount. One feature of that scale is, of course, that its zero point and unit are arbitrary. In isolated experiments there must be some way of specifying the location and unit for θ. The usual procedure is to fix the mean at 0 and the standard deviation at 1. Waller appears to claim that this is not enough: When the original θ distribution is badly skewed, there appears to be severe bias in the estimation of the item parameters.

Part of the problem is readily solved with a scale adjustment. Waller's original distribution of θ had a mean of −.83 and a standard deviation of 1.18. The original values of the item parameters had average difficulty of 0 and average discriminability of 1.0 *on this scale*. Yet, the LOGOG computer program sets the mean of the ability distribution to 0 and the standard deviation to 1 and reports estimates of the item parameters *on that scale*. If Waller had transformed the original item parameters to correspond with a standardized θ scale, they would have had an average difficulty of .83 and an average discriminability of 1.18 (assuming arithmetic averages). In fact, Waller observed an average difficulty of .90 and an average discriminability of 1.27. Thus, most of the difference seems to be artificial and to be due to a scale shift.

How much of the remaining difference is bias and how much is sampling error? Since only one sample of 480 pseudo-cases and 45 pseudo-items were tried, there is no way to tell: One sample does not make a monte carlo study. Swaminathan and Gifford did a similar study with a *negatively* skewed distribution (Waller's was *positively* skewed) and obtained difficulties and discriminabilities that were too high. This would seem to be consistent for the discriminabilities and to be inconsistent for the difficulties.

If the θ scale is important, then its metric is important, in which case why did Hambleton and Cook use rank-order correlations to evaluate correspondence of θ and $\hat{\theta}$? Product-moment correlation would seem to be the obvious choice. They claimed that the scale of θ is arbitrary. The origin and unit are arbitrary, but the metric is not. If the metric were arbitrary, of what value is latent trait theory?

Some investigators believe that the $\hat{\theta}$ scale, at least the maximum likelihood $\hat{\theta}$ scale, is unfit for linear statistical methods. If so, then latent trait theorists have an inferior product. The problem is at the extremes, where ability estimates can have huge standard errors. Lord advocates transforming back to the true score scale, which I thought was what we were escaping, whereas Novick advocates (I suppose) Bayesian estimation. Bayesian estimates have no problem, because an infinite value of $\hat{\theta}$ has an infinitesimal a priori probability, so that only an infinitely perverse examinee will give any trouble. There are other possibilities. Why not refuse to give a score to an extreme person? Of course, an adaptive test with an adequate supply of items would, at least in principle, be in a much better position. Since in such a test the item difficulties match the person's ability, this "end effect" should be a much smaller problem. One way or another, though, this end problem needs to be resolved. There is no future for test scores that are unsuited to linear statistical methods.

Hambleton and Cook and Swaminathan and Gifford have studied the properties of estimates of ability and the item parameters as functions of the number of examinees, the number of items, and the true distribution of ability. They used "constructed" data and the monte carlo approach. Note carefully that each tabular entry is based on only one sample data matrix. Although the entries are averages over items and persons, in a real sense, each of the entries represents one sample point. Thus, individual entries are not to be relied upon; only general trends should be interpreted.

Hambleton and Cook evaluated the fit of the 1-, 2-, and 3-parameter models to data from each of these three models with a uniform distribution of ability. They also compared the lower and upper halves of the ability distributions. With only 20 items, the 1- and 2-parameter models were poorer in the lower half than in the upper half of the ability distribution. When the entire ability distribution was analyzed, 40 items were slightly better than 20, and ability was better estimated when there was no guessing. All three models fit a given data matrix almost equally well, but apparently this is a very good set of "items." (Roughly the same pattern of results was found for a normal distribution of ability, but all values were smaller.)

Much more interesting are their results concerning the standard errors of estimate as a function of ability. Clearly, a 10-item test gives unsatisfactory standard errors; a 20-item test is not very good for low abilities; but an 80-item test gives nearly constant standard errors for abilities in the range -2.0 to +2.0. The typical great increase of standard error occurs for more extreme scores. (This is the unfortunate problem with the ability metric that was mentioned above.)

I would like to know not only how well each method did relative to the true values but also how the methods compared with each other. What are the correlations of $\hat{\theta}$ with θ for the 1-, 2-, and 3-parameter models? Almost certainly, they were extremely high.

Swaminathan and Gifford compared two estimation procedures and found LOGIST to be superior. They also showed that as both the number of items and the number of persons increased jointly, the LOGIST parameter estimates approached the true values without bias, indicating empirical consistency. This demonstration is heartening but would be more convincing with more data sets, i.e., more replications.

Swaminathan and Gifford showed that Urry's method had trouble estimating the guessing parameter. It would be interesting to know if the other problems with the method were related to this flaw. Why not estimate a single guessing parameter for all items, or at least for all items of a given type? Or, if there are few enough items, why not set $\underline{c} = .20$ or .25, or whatever seems empirically reasonable, and only estimate the other two parameters for each item?

Note that Hambleton and Cook and Swaminathan and Gifford asked how large N and n should be. By contrast, Lord asked which procedure should be used if N is small. He reasoned, and found, that for an N small enough (roughly 100) the 1-parameter model was actually superior. Given the recent work on equal weights in regression, that result must inevitably be so. Empirically determined weights are uncertain with small N.

Hambleton and Cook claimed that samples of 200 persons and 20 items are satisfactory for some applications of latent trait theory. It is very important that their conclusions be noted carefully and that that claim not be over-generalized. Certainly, when the model fits the data, the item parameters can be adequately estimated. The ability parameters can also be estimated, but the standard errors are large, and the extreme cases are still a problem. A standard 20-item test will not give very reliable results, no matter what theory is used. And 80 items and a great many examinees would be very much better than 20 items and only 200 examinees.

How realistic are these studies? First, all of them used data constructed by monte carlo methods. Lord based his theoretical item parameters on those from an actual data set--a 30-item subset from a 50-item vocabulary test. The range of item discrimination indices was .4 to 1.8, with quartiles of .55, .83, and 1.35. Swaminathan and Gifford used a uniform distribution in the range .6 to 2.0, a distinctly better set of items. Hambleton and Cook used two ranges--.5 to 1.74 and .81 to 1.43--much like Lord's set. All of these are good items, with excellent discrimination. Also, the items were constructed to be unidimensional. What happens with items of more ordinary discriminability and with some secondary group factors?

Secondly, how often will the model be applied when item parameters are unknown? Is it not at least as likely that calibrated items will be available from which only ability needs to be estimated? Suppose a few uncalibrated items are being pretested; item parameters are to be estimated in the context of the calibrated items and the estimated ability scores. Most especially, how does this kind of conditional estimation proceed in a computerized adaptive testing environment. This seems a good place to apply sequential Bayesian procedures. Finally, what happens with real data? Simulation studies have their place, but much more is to be learned with real data.

SESSION 9:

LONGITUDINAL MEASUREMENT WITH LATENT TRAIT MODELS

SOME LATENT TRAIT MODELS FOR MEASURING CHANGE IN QUALITATIVE OBSERVATIONS

GERHARD FISCHER
UNIVERSITY OF VIENNA

THE MENTAL GROWTH CURVE RE-EXAMINED

R. DARRELL BOCK
UNIVERSITY OF CHICAGO

LATENT STRUCTURE ESTIMATION FOR ASSESSING GAIN IN ABILITY

LALITHA SANATHANAN
ARGONNE NATIONAL LABORATORY

SOME LATENT TRAIT MODELS FOR MEASURING CHANGE IN QUALITATIVE OBSERVATIONS

GERHARD FISCHER
UNIVERSITY OF VIENNA

All too often, thinking and formation of concepts in behavioral science have been misled by readily available statistical methods, especially of the multivariate variety. A typical example of how theorizing in psychology can be led astray by statistical methods is the obsolete dispute on the percentage of genetically versus environmentally determined intelligence. In this field there has been an unwavering attempt to apply models of variance decomposition, which had been developed for breeding experiments in stock-farming and which are appropriate for that purpose; however, these methods are not suited for yielding scientific insights into the genetic and environmental factors of human intelligence. On the other hand, there has been a failure to develop adequate methods for answering the question, What is the effect of specified types of socioeconomic environment on the development of human intelligence?

However, it is methodology that must be adjusted to the theoretical concepts and problems in applied behavioral science, rather than the reverse. This is illustrated by an example from communication research: In 1971 a basic problem of market and opinion research was posed, namely, What is the effect of an insertion in different media, such as television, radio, or newspapers? For the practical purpose of optimizing a campaign with a limited budget, a simple answer to the question was needed, e.g., "An insertion in television is three times as effective as a comparable insertion in a local radio program." In addition, it seemed that it was chiefly the methods currently used in communication research that were responsible for the lack of generalizable results on communication effects. The problem was as follows: Suppose it were possible to describe the effectiveness of each medium by just one quantitative parameter; suppose further that each interviewed person could be characterized by certain attitude parameters pertaining to the topic of the campaign and by the subject's individual amount of consumption of each medium. What kind of probabilistic model would then give a straightforward answer to the simple question, What is the effect of medium $\underline{j}$ relative to the effect of medium $\underline{k}$?

At the same time, for theoretical as well as for practical reasons, there was an attempt to comply with the principle of specific objectivity, as introduced by Rasch (1967, 1972): The comparison of the effect parameters of two media should depend on these two parameters only and should be independent of any irrelevant factors, such as the parameters characterizing the initial attitudes of the respondents. In other words, the result should be independent of the sample of respondents.

These considerations resulted in a family of logistic models that are closely related to the well-known Rasch (1960, 1966) models but also show some marked distinctions. Unfortunately, the models have been applied to assessing effects of mass communication only once; but many problems in clinical and educational psychology are of similar structure, the media being replaced by therapeutic or educational treatments. A considerable number of applications in these fields have been undertaken in the last five years, and the theoretical and methodological bases of the models have been further strengthened.

As can easily be seen, the question regarding the effects of mass communication is nothing but a special case of the question of change under the influence of some sort of treatment. Therefore, the models referred to are of considerably broad interest. Their distinction from more conventional approaches to measurement of change is that the data are regarded as what the observations, in fact, mostly are: qualitative variables. In this paper it will by no means be attempted to scale or to quantify the data in order to make the classical statistical methods applicable. Quite the contrary, the observations will be explained as realizations of qualitative random variables, which are, however, governed by quantitative latent parameters. Change is defined as a change in these latent parameters.

Models for Qualitative Data

There are a variety of such models, differing as to the restrictiveness of their assumptions, the kind of results deducible, and the required types of data. The most important models are:

1. The dichotomous linear logistic test model (LLTM), which was originally devised for analyzing the complexity of intelligence test items in terms of cognitive operations involved, but is also useful for measuring change in unidimensional latent variables or for certain experimental designs with more than two points of time. Since the formalism of this model is rather complicated, it will not be dealt with here (see Fischer, 1973, 1974a, 1974b, 1977a).

2. The dichotomous linear logistic model with relaxed assumptions (LLRA), which emphasizes the relaxation of assumptions as compared with the usual latent trait models, since no unidimensionality of the criterion variables or items is assumed. It has proven a very useful tool for assessing change in a variety of different situations and will be described in this paper.

3. The polychotomous extension of the LLTM, for which applications are lacking. Since this paper will not dwell on purely theoretical developments that have not as yet stood the test of practical application, this model will merely be mentioned (see Fischer, 1974a, 1974b, 1977c).

4. The polychotomous generalization of the LLRA, offering quite interesting possibilities of application and empirical hypotheses testing, which will be mentioned below.

The Dichotomous LLRA

The Model

"Dichotomous" means that the observed criterion variables, which may be test items or clinical symptoms or any other kind of behavior, are binary variables. It is assumed that before and after treatment a number of k such criterion variables are observed on each subject. Then, the model is defined by the following equations:

$$P(+|v,i,t_1) = \frac{\exp(\xi_{vi})}{1+\exp(\xi_{vi})} \qquad [1]$$

$v = 1,\ldots,n$ (subjects)

$i = 1,\ldots,k$ (criteria, items),

$$P(+|v,i,t_2) = \frac{\exp(\xi_{vi}+\delta_v)}{1+\exp(\xi_{vi}+\delta_v)} \qquad [2]$$

$$\delta_v = \sum_j q_{vj}\eta_j + \sum_l \sum_j q_{vj} q_{vl} \rho_{jl} + \tau \qquad [3]$$

Thereby, $P(+|v,i,t_1)$ denotes the probability that subject v gives response "+" in criterion i at time t_1 (before treatment) and that $P(+|v,i,t_2)$ is the analogous probability for time t_2 (after treatment). The probability $P(+|v,i,t_1)$ depends solely on one parameter, ξ_{vi}. For example, let criterion i be a certain symptom of fear in clinical patients, then ξ_{vi} is the latent anxiety of subject v behind that symptom. Thus, the state of subject v at time t_1 is characterized by a vectorial parameter $\xi_v = (\xi_{v1},\ldots,\xi_{vk})$, in other words, by a set of k traits associated with the k criterion variables.

Note that the model makes no assumptions whatsoever about interdependencies or dimensionality of these traits; in particular, unidimensionality of the criteria or items is not assumed, as would be the case with the Rasch (1960, 1966) or Birnbaum (1968) models. Hence, the LLRA is maximally flexible regarding the characterization of the subjects. For example, it may well be that $\xi_{vi} < \xi_{vj}$, but that $\xi_{wi} > \xi_{wj}$.

The characterization of the subjects at time t_2 is, in principle, analogous; it is, however, restricted by the assumption that change in each subject can be described by a single parameter δ_v, which according to Equation 3 is a linear function of the effects η_j of the given treatments (main effects), of their interactions ρ_{ij}, and of a trend-parameter τ comprising all the causes of change that are unrelated to the treatments. The constants q_{vj} are measures of the dose of treatment j as applied to subject v.

The most important properties of the model are:

1. Given appropriate data, the effect parameters η_j, the interactions ρ_{ij}, and τ can be estimated independently of the true values of the parameters ξ_{vi}; the latter need not be known and are not estimated from the data, either. This means that any proposition referring to the comparison of two treatments i and j is completely independent from the sample of subjects (specific objectivity).

2. The parameter estimates are a ratio scale, so that it is possible to arrive at statements such as "treatment i is twice as effective as treatment j."

3. It is possible to test the significance of single parameters and to test almost any conceivable meaningful composite hypothesis on the parameters by means of likelihood-ratio tests.

The formal properties of the model have been studied by Fischer (1972, 1974a, 1974b, 1976, 1977a, 1977c; see also Fischer & Rop, in prep.).

The sheer enumeration of the model properties does not sufficiently reveal the full scope of the possibilities implied by these properties. An illustrative example will therefore be in order.

Sample Application

Research questions. Rop (1977) investigated the effects of three preschool educational programs (Early Reading, Logical Thinking, and Verbal Enrichment) on the cognitive development of kindergarten children. To assess change, a battery of 64 items was given before and after the treatment period; a control group attended kindergarten but did not participate in the programs. Three primary questions were to be answered:

1. Is it possible to furnish proof that the programs accelerate cognitive development?

2. What is the generality of the effect of each of the programs, e.g., is there an effect of verbal training also in the nonverbal area?

3. What do the socially and educationally disadvantaged children gain in comparison to middle-class children, i.e., is early intervention a means of overcoming the deficiencies resulting from less privileged environments?

The first question, which is the easiest one, was answered in the affirmative by testing the null hypothesis that the effect parameters are zero.

The second question is far more complex: Equation 3 asserts that the effect of each treatment can be measured by just one parameter δ_v per person, irrespective of the item i. Hence, if it were true that verbal enrichment had little or no effect on the nonverbal abilities (which were tested by one set of nonverbal items in the test), the model could not have been true for all 64 items. Hence, the model plays the role of a H_0 against the H_1 of differential effects of the treatments in certain subgroups of items, i.e., the criterion variables. (There is a far-reaching analogy between this model and the well-known analysis of variance for quantitative data: In analysis of variance as well, one begins with the global H_0 that all means are equal.)

Results. In Rop's study the H_0 of uniform effects of the treatments on all the ability domains represented by the items had to be rejected. As Table 1 shows, the 64 items had to be broken down into three subsets (naming of objects, actions, and attributes; verbal abilities, such as verbal fluency, enunciation, and appropriate usage of language; and nonverbal abilities). Each of the three programs had a differential effect within each of the three domains. However, the results in Table 1 show the findings of the study in a maximally generalized form: It is an essential feature of the model that it identifies the maximal subsets of criterion variables with uniform treatment effects. This is a consequence of the principle of specific objectivity, viz., that the estimates of effect parameters do not depend on any irrelevant factors, such as subjects or items, as long as the model holds. In other words, only the minimum number of moderator variables that are absolutely necessary to explain the data are considered.

Table 1
Effects of the Training Programs per Time Unit (1,000 minutes) and the Trend for Naming, Verbal Intelligence, and Nonverbal Item Groups

	Item Group		
Treatment	Naming	Verbal Intelligence	Nonverbal
Reading	.37*	.15	-.04
Thinking	.51*	.16*	.25*
Verbal	.49*	.37*	.31*
Trend	.84*	.88*	.32*

*Statistically significant at $p \leq .01$ (Adapted from Rop, 1977).

Rop's third question is the most intriguing one. If conventional methods of data analysis had been applied, it would have been expected that environmentally privileged children with a higher level of cognitive development, and hence with better performance at t_1, would not have increased their level of performance as much as the children with poor achievement. Such methodological artifacts are known under the names "physicalism-subjectivism-dilemma," "base-rate problem," or the like (see Bereiter, 1967; Lord, 1967). The LLTM, on the contrary, asserts that the effect parameters do not depend on the subject parameters ξ_{vi}; in other words, if the effect of treatments were really the same for all children, the effect parameters estimated from groups of children with different ability levels should also be equal except for random error. This is again a direct consequence of the principle of specific objectivity. In Rop's study it was found, in fact, that treatment effects were independent of the initial level of cognitive development and therefore that the preschool programs were not appropriate for bridging the gap between privileged and underprivileged children.

Measuring Change with the LLRA Model

It is obvious that the properties of the LLRA model are quite advantageous, having encouraged a variety of applications. However, a better theoretical and epistemological foundation of this methodological approach seems called for, and an answer was sought to the following question: If assessment of change is to be specifically objective, what is implied with respect to the formal structure of the model? A prerequisite for dealing with this question is to formalize the problem of measurement of change in a sufficiently general way.

The Model

Change is detected by exposing subjects to a set of observational conditions such as the test items, observation of symptoms, or registration of any other kind of criterion variables. Let the behavioral disposition or state of the subject at time $\underline{t}_1$ be described by a set of $\underline{k}$ parameters $\rho_{v1,1}(\xi_{v1}),\ldots,\rho_{vk,1}(\xi_{vk})$, so that $\rho_{vi,1}$ is associated with the criterion variable $\underline{i}$ and describes fully the latent behavioral disposition of subject $\underline{v}$ with respect to this variable. In the same way, let the state of subject $\underline{v}$ at time $\underline{t}_2$ be described by the set of parameters $\rho_{v1,2}(\xi_{v1}, \delta_v),\ldots,\rho_{vk,2}(\xi_{vk}, \delta_v)$, whereby δ_v is a scalar parameter representing change. Nothing is assumed concerning the functional concatenation between ξ_{vi} and δ_v; it is only assumed that the reaction tendency $\rho_{vi,2}$ at time $\underline{t}_2$ is a function of the latent trait ξ_{vi} at time $\underline{t}_1$ and the change parameter δ_v. Since the objective is to assess change, the existence of a function $U(\rho_{v1,1},\ldots,\rho_{vk,1};\rho_{v1,2},\ldots,\rho_{vk,2})$, which can be solved for δ_v, will further be assumed. In other words, U should be a function of δ_v alone: $U = V(\delta_v)$. It is a consequence of the principle of specific objectivity that U must be independent of the latent trait parameters ξ_{vi} and of the sample of observational situations chosen for assessing change.

On the basis of this formalization of measurement of change, the following theorem can be proven:

Theorem 1. Let $\partial\rho_{vi,1}/\partial\xi_{vi} > 0$ everywhere; let U be differentiable with respect to $\rho_{vi,1}$ and $\rho_{vi,2}$; $\rho_{vi,1}$ with respect to ξ_{vi}; and $\rho_{vi,2}$ with respect to ξ_{vi} and δ_v for $\underline{i} = 1,\ldots,\underline{k}$. Further, let U be a function $U(\rho_{v1,1},\ldots,\rho_{vk,1};\rho_{v1,2},\ldots,\rho_{vk,2}) = V(\delta_v)$, which is independent of ξ_{vi}, $\underline{i} = 1,\ldots,\underline{k}$. Then, there exist monotone transformations of all the parameters, so that after transformation, $\rho_{vi,2} = \rho_{vi,1} + \delta_v$. If, in addition, the observations are assumed to be realizations of Bernoulli variables, then, except for scale transformations, Equations 1 and 2 must hold. (The proof of this theorem is rather complicated; see Fischer, 1977b; Fischer & Rop, in prep.; Rasch, 1972). The meaning of the theorem is this: If there are dichotomous observations at two points of time and if it is desired to assess change in a specifically objective manner (i.e., if the result should not depend on the sample of person-parameters ξ_{vi}), then the model must be essentially of the LLRA type. Of course, some scale transformations may be applied on the parameter dimensions, entailing formal changes of the model, but any model and empirical result obtained in this manner would be completely equivalent to what is obtained by means of the LLRA. Hence, there is no point in transforming the parameters and thereby departing from the specifically simple structure of the LLRA.

Estimating Model Parameters

Since this theorem legitimates the LLRA as theoretically well founded, a short discourse on the technical problems of parameter estimation and hypotheses testing is called for. To simplify matters, Equation 3 can be rewritten as

$$\delta_v = \sum_j q_{vj}\eta_j \quad . \qquad [4]$$

This can be done because Equation 3 is linear in all the parameters; hence, parameters η_j and matrix $\underset{\sim}{Q} = [(q_{vj})]$ need to be redefined appropriately.

Theorem 2. Let $A_1 = [(a_{vi,1})]$ be the item-score matrix (with elements 1 if the response was "+" and 0 if "-") for time t_1 and $A_2 = [(a_{vi,2})]$ for time t_2, $v = 1,\ldots,n$ and $i = 1,\ldots,k$. The conditional maximum likelihood estimates $\hat{\eta}_j$ of the effect parameters η_j are given by the equations

$$\sum_v \sum_i q_{vj}\left\{a_{vi,2} - \frac{\sum_{t=u}^{w} \exp\left(t\sum_j q_{vj}\hat{\eta}_j\right)t}{\sum_{t=u}^{w} \exp\left(t\sum_j q_{vj}\hat{\eta}_j\right)}\right\} = 0 \ , \qquad [5]$$

$v=1,\ldots,n$; $i=1,\ldots,k$; $j=1,\ldots,m$;
$u(v,i) = w(v,i) = 0$ if $a_{vi,1}+a_{vi,2} = 0$;
$u(v,i) = w(v,i) = 1$ if $a_{vi,1}+a_{vi,2} = 2$; and
$u(v,i) = 0$, $w(v,i) = 1$ if $a_{vi,1}+a_{vi,2} = 1$.

The estimation Equations 5 have a finite solution $\hat{\eta} > 0$ if for $j = 1,\ldots,m$ holds

$$\sum_v \sum_i q_{vj} a_{vi,1}(1 - a_{vi,2}) > 0 \qquad [6]$$

and

$$\sum_v \sum_i q_{vj}(1 - a_{vi,1})a_{vi,2} > 0$$

for $j = 1,\ldots,m$;

the solution is unique if the rank of $Q = [(q_{vj})]$ equals m, and it is at a maximum of the likelihood function.

The estimation Equations 5, and the corresponding second-order partial derivatives that are needed for applying the Newton-Raphson procedure, were given by Fischer (1972, 1974a, 1974b, 1977a, 1977c; for the complete proof of Theorem 2, see Fischer and Rop, in prep.)

This theorem is not only useful for determining the existence and uniqueness of the solution but it also implies that the parameter estimates lie on a ratio scale with a unit determined by the time interval (t_1, t_2) together with the chosen unit of measurement of dosage q_{vj}.

Hypothesis Testing

In practical applications, the estimation of the parameters is only a first step. More important is the test of hypotheses on the parameters. Such tests can be carried out by means of the likelihood ratio principle.

Let $\hat{\eta} = (\hat{\eta}_1,\ldots,\hat{\eta}_m)$ be the estimates of effect parameters under hypothesis H_1 (alternative hypothesis) and let $L(H_1)$ be the maximized conditional likelihood of the data under H_1,

$$L(H_1) = \prod_v \prod_i L_{vi}, \text{ with} \qquad [7]$$

$$L_{vi} = \begin{cases} \dfrac{\exp(a_{vi,2} \sum_j q_{vj}\hat{\eta}_j)}{1 + \exp(\sum_j q_{vj}\hat{\eta}_j)} & \text{if } a_{vi,1} + a_{vi,2} = 1 , \\ 1 \text{ if } a_{vi,1} + a_{vi,2} = 0 \text{ or } = 2 . \end{cases}$$

Further, let H_0 be a null hypothesis consisting of the restrictions $\eta_j = \mu_j(\beta_1,\ldots,\beta_{m'})$ with $m' < m$, whereby the matrix of partial derivatives $\partial\mu_i/\partial\beta_l$ has rank m'. Finally, let $L(H_0)$ be the likelihood of the data under H_0, whereby the maximum likelihood estimates $\hat{\eta}^*_j = \mu_j(\hat{\beta}_1,\ldots,\hat{\beta}_{m'})$ are inserted in Equation 6 instead of $\hat{\eta}_j$. Then, under H_0,

$$-2ln\lambda = -2ln\{L(H_0) - L(H_1)\} \qquad [8]$$

is asymptotically chi-square-distributed with $df = m - m'$. It can easily be shown that most hypotheses relevant in practical applications can be formulated as restrictions $\eta_j = \mu_j(\beta_1,\ldots,\beta_{m'})$ and hence can be tested by means of this likelihood ratio test. As long as the restrictions are linear contrasts, estimation Equations 5 can be used for estimating the parameters under H_0 as well;

otherwise, the estimation equations require a minor adaptation, which need not be discussed here.

It is a basic feature of the model that, in a formal respect, it makes no difference whether a new set of subjects or additional criteria are added to the given observations. Therefore, differences of treatment effects between subsets of subjects and between subsets of criteria lend themselves to exactly the same kind of test. Some examples of hypotheses typically tested in the applications are the following:

1. All interactions are zero ($\rho_{ij} = 0$, i, $j = 1,\ldots,m$).
2. Some treatments are equally effective (e.g., $\eta_j = \eta_1$).
3. Some treatments are ineffective (e.g., $\eta_j = 0$).
4. The trend effect is zero ($\tau = 0$).
5. The effect of treatments and/or the trend effect is equal for different subgroups of subjects or in different subgroups of criteria ($\hat{\eta}_j(\mathrm{I}) = \hat{\eta}_j(\mathrm{II})$ for Groups I and II).

In principle, the tests are logically analogous to hypothesis testing in linear analysis of variance and to testing linear contrasts between groups of mean values.

An interesting special case arises when testing the dose-response relationship: Leaving aside the question of interactions, the model Equation 3 presupposes that the effect of treatment is proportional to the dose. However, general experience indicates that in some cases a treatment is completely ineffective below a certain minimal dose and that above a certain amount of treatment satiation occurs. It is therefore important that the hypothesis of linearity, Equation 3, is tested against an unspecified nonlinear dose-response curve. This can be done by means of the following parameterization: Suppose that dose is no continuous variable but assumes certain discrete values $u_1,\ldots,u_s$. Then, it is possible to assign one parameter $\eta_{j1},\ldots,\eta_{js}$ to each of these doses. To embody this set of new parameters into the model, let $\underline{b}_{vj} = (\underline{b}_{vj,1},\ldots,\underline{b}_{vj,s})$ be a selection vector with elements

$$b_{vj,t} = \begin{cases} 1 \text{ if subject } v \text{ has obtained dose } u_t \text{ in treatment } j \text{, and} \\ 0 \text{ otherwise.} \end{cases} \qquad [9]$$

The selection vector for each combination of subject v × treatment j consists of 0's except for one element, which is equal to 1 and indicates the dose obtained in the respective treatment. The model Equation 4 then becomes

$$\delta_v = \sum_j \sum_t b_{vj,t}\eta_{jt}. \qquad [10]$$

Now, consider Equation 4 as H_0 and Equation 10 as H_1, allowing a likelihood ratio test of the linearity hypothesis.

Applications of the LLRA in Measuring Change

Rop's study on the effects of preschool education has already been mentioned above: The programs (e.g., logical training) did not just affect the narrow domain of the functions trained but also influnced the other intellectual factors, if less markedly. This problem of transfer of cognitive operations has been discussed by Zeman (1976), who investigated the effects of early training in elementary set theory; she proved a substantial transfer of the operations acquired from material used in the learning phase to other materials. This finding implies that, as was hoped, this specific preschool education is in fact a rather general vehicle for promoting cognitive development.

An interesting application of the LLRA to clinical psychology stems from Heckl (1976), who investigated the effects of three forms of speech therapy in children with speech disorders. Contrary to expectation, all three therapies proved to be equally effective. The interpretation was that the effect apparently was brought about by the intensive devotion of the therapist to the handicapped children and by the reinforcement given to their verbal productions--relatively independent of the content of the prescribed exercises. Heckl's study is one of the few where the linearity of the dose-response curve was empirically tested: A substantial difference in effect between children with fortnightly therapeutic sessions and children with one session per week was observed; the further benefit of two sessions per week, however, was comparatively small. Apparently, satiation occurred in the latter case. As in Rop's study, the effect parameters were (at least approximately) constant over different groups of children, i.e., independent of age, sex, and only partially dependent on the degree of initial speech impediment.

Another study in the domain of clinical psychology is that of Glatz (1977), who investigated the effects of behavior therapy on the eating performance of mentally retarded children. A special feature of this study is that observations were made at eight points in time, yielding a behavioral sequence for each child. Glatz used the LLRA for comparing two successive points of time each; strictly speaking, however, another type of linear logistic model, the LLTM, would have been more appropriate. A reanalysis of the data is underway (Fischer & Rop, in prep.).

There have been several additional applications of the model. Vodopiutz (1977) studied the effects of certain training units on complex movements in gymnastic education; Pendl (1976), the effects of a language laboratory on teaching a foreign language (English) in high school; Rella (1976), the results of driver improvement training in anticipating dangerous traffic situations; Platzer (1978), the effects of technical playing materials on the development of mechanical-technical understanding; Witek (1979), the effects of a group-dynamic sensitivity training for business executives; and Zimprich (1979), the effects of psychotherapy, given in addition to chemotherapy, to patients of an internal department of a children's hospital.

The Polychotomous LLRA

The Model

Although quite often the data are readily reducible to dichotomous variables, in many cases such a reduction either is not possible or makes little sense. In spite of this, designs with polychotomous data have not received enough attention in the literature owing to the lack of suitable methodology. Already in the early papers on linear logistic models for measuring change (Fischer, 1972, 1974a, 1974b, 1977c), the possibility of generalizing the LLRA to polychotomous data has been recognized and the necessary estimation equations have been derived. Without going into technical details, the essentials of the parameterization will be presented here.

Suppose that k polychotomous variables, each of which may assume one of r qualitative or quantitative realizations, are the basis for assessment of change. A generalization of the model Equations 1 to 3 is then

$$P\left(A_{vi}^{(h)} = 1 | v,i,t_1\right) = \frac{\exp\left(\xi_{vi}^{(h)}\right)}{\sum_t \exp\left(\xi_{vi}^{(t)}\right)} , \qquad [11]$$

$$P\left(A_{vi}^{(h)} = 1 | v,i,t_2\right) = \frac{\exp\left(\xi_{vi}^{(h)} + \delta_v^{(h)}\right)}{\sum_t \exp\left(\xi_{vi}^{(t)} + \delta_v^{(t)}\right)} , \qquad [12]$$

$$\delta_v^{(h)} = \sum_j q_{vj} \eta_j^{(h)} + \tau^{(h)} , \qquad [13]$$

$$v = 1,\ldots,n; \quad i = 1,\ldots,k;$$

$$j = 1,\ldots,m; \quad t,h = 1,\ldots,r.$$

Thereby, $A_{vi} = (A_{vi}^{(1)},\ldots,A_{vi}^{(r)})$ is an indicator vector-variable with realizations $a_{vi}^{(h)} = 1$ if subjects v's reaction on criterion i was in category h, and $a_{vi}^{(h)} = 0$ otherwise. The state of each subject at t_1 is characterized by a matrix of parameters $\xi_{vi}^{(h)}$ and "change" is described by a vectorial parameter $\delta_v = (\delta_v^{(1)},\ldots,\delta_v^{(r)})$; element $\delta_v^{(h)}$ measures the effect of the treatments with respect to reaction category h. Analogously, the effect of each treatment is described by a vectorial parameter $\eta_j = (\eta_j^{(1)},\ldots,\eta_j^{(r)})$, its elements being assocated with the specific effect of treatment j with respect to response category h. To be more concrete, the behavior categories of a depressive patient could be, for example, *agitated*, *withdrawn*, and *normal*; a certain psychiatric treatment could then have a very strong effect of reducing agitation and increasing withdrawal without, however, necessarily increasing the rate of normal behavior.

Several such qualitative categories may as well express different levels of an underlying latent dimension, i.e., different degrees of *one* behavioral tendency. A typical example would be the categories *very content*, *rather content*, *rather not content*, *not at all content*, reflecting degrees of satisfaction (e.g., with a job). The case of unidimensionality of the response catgories with respect to the treatments is then formalized as follows:

$$\eta_j^{(h)} = \phi^{(h)}\eta_j \qquad (j = 1,\ldots,k;\ h = 1,\ldots,r)\ . \qquad [14]$$

Equation 14 has been called the reduction conditions; of course, it is a purely empirical matter whether they hold or not. If they hold, the matrix of parameters $\eta_j^{(h)}$ is of rank 1. The parameters $\phi^{(h)}$ are called the category weights.

As in the dichotomous LLRA, the effect parameters and the trend effects can be estimated empirically, independent of the person parameters $\xi_{vi}^{(h)}$, which characterize the state of the sample at t_1. Furthermore, hypotheses are testable by means of likelihood ratio tests. One reservation, however, must be made regarding the reduction conditions: When the parameters are estimated under assumption of Equation 14, the solutions of the estimation equations are not necessarily unique.

Applications of the Polychotomous LLRA

The numerical computations for estimating parameters in the polychotomous case are much more complex than in the dichotomous case, and some theoretical questions need further investigation (as, for example, uniqueness of the solution in case of the reduction conditions). In addition, the amount of data required is much larger than in the dichotomous LLRA. For these reasons, only a few empirical applications have been realized so far. Nevertheless, the polychotomous LLRA is a potentially powerful instrument for assessing change, as will be illustrated by the following two empirical studies.

Hammer (1978) investigated the cognitive and attitudinal effects of a multi-media presentation dealing with forms of human settlement, problems of big cities, and ecology. The presentation was viewed by one sample of high-school children, whereas another sample received instruction on the same topics from a teacher. The cognitive effects of both methods of instruction were measured by a questionnaire with the three response categories *correct*, *partially correct*, and *incorrect*; the attitudinal/emotional effects were evaluated by another questionnaire with categories *positive*, *neutral*, *negative*, and *don't know*. As expected, the multi-media presentation proved to be generally more effective than the teacher, especially so with respect to the domain of attitudinal and emotional change; the teacher was able to impart knowledge rather than to influence attitudes or to appeal to emotions.

The second example of an application of the polychotomous LLRA returns to the problem of measuring effects of mass communication mentioned earlier: Kropiunigg (1979) carried out a field study on a topical problem of social and political interest in Austria on the reform of penal law in 1975. In Styria, one of the nine provinces of Austria, an informational campaign on this topic was promoted by the *Regionalprogramm Studio Steiermark* (radio) and by the *Kleine Zeitung Graz* (newspaper), whereby problems of probation and resocialization of convicts were dealt with.

Before and after the campaign, representative samples of the population were interviewed (t_1: $n = 550$; t_2: $n = 640$). The questionnaire comprised items referring to three attitudinal domains and one set of items for assessing familiarity with relevant facts. Since the subjects interviewed at t_1 and t_2 (unlike the case of the standard LLRA) were not the same, a modified version of the model for independent samples had to be used (see Fischer, 1972, 1974a, 1974b, 1977c).

This study differed from those of the other above-mentioned investigations in one essential respect: It was not possible to obtain generalizable propositions with respect to the effects of the media. The results rather supported the standard conjecture of communication theory: that effects of communications are strongly determined by a number of moderator variables (e.g., socioeconomic factors). Only the result that the radio programs were more effective than the respective articles of the daily newspaper was of some generality. Those segments of the population characterized by high contact frequency with the radio programs in question showed satiation regarding the information on the issue; an increase of density of the pertinent information in the newspaper, on the other hand, would still have increased the effects of the campaign. A somewhat unexpected finding was the relatively limited acceptance of the promoted ideas by women and by religious people, whereas supporters of the (governing) socialist party showed significantly above-average understanding.

The principal goal of giving a simple characterization of each medium by a few effect parameters $\eta_j^{(h)}$--which had originally led to the development of the LLRA and other linear logistic models--was not reached in this empirical study. Perhaps the epistemological basis of these considerations is not appropriate for the complex problem of social science. But the theoretical developments and the applications in other fields, as mentioned above, indicate that it was worthwhile to venture models that derive very simplified and generalized results from complex bodies of qualitative data.

REFERENCES

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: The University of Wisconsin Press, 1967.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of men-*

tal test scores. Reading, MA: Addison-Wesley, 1968.

Fischer, G. H. A measurement model for the effect of mass media. Acta Psychologica, 1972, 36, 207-220.

Fischer, G. H. The linear logistic model as an instrument in educational research. Acta Psychologica, 1973, 37, 359-374.

Fischer, G. H. Einführung in die theorie psychologischer tests. Bern: Huber, 1974. (a)

Fischer, G. H. Lineare logistische modelle zur beschreibung von einstellungs-- und verhaltensänderungen unter dem einfluss von massenkommunikationen. In W. F. Kempf (Ed.), Probabilistische modelle in der sozialpsychologie. Bern: Huber, 1974. (b)

Fischer, G. H. Some probabilistic models for measuring change. In D. de Gruijter & L. van der Kamp (Eds.), Advances in psychological and educational measurement. New York: Wiley, 1976.

Fischer, G. H. Linear logistic test models. In H. Spada & W. F. Kempf (Eds.), Structural models of thinking and learning. Bern: Huber, 1977. (a)

Fischer, G. H. Some implications of specific ojectivity for the measurement of change (Research Bulletin No. 21). Vienna: University of Vienna, Institute of Psychology, 1977. (b)

Fischer, G. H. Some probabilistic models for the description of attitudinal and behavioral changes under the influence of mass communication . In W. F. Kempf & B. Repp (Eds.), Mathematical models for social pyschology. Bern: Huber, 1977. (c)

Fischer, G. H., & Rop, I. Latent trait models for measuring change. Manuscript in preparation.

Glatz, E.-M. Die wirksamkeit eines verhaltenstherapeutischen ess--Tranings bei geistig retardierten kindern. Unpublished doctoral dissertation, University of Vienna, 1977.

Hammer, H. Informationsgewinn und motivationseffekt einer tonbildschau und eines verbalen lehrervortrages. Unpublished doctoral dissertation, University of Vienna, 1978.

Heckl, U. Therapieerfolge bei der behandlung sprachgestorter kinder. Unpublished doctoral dissertation, University of Vienna, 1979.

Lord, F. M Elementary models for measuring change. In C. W. Harris (Ed.), Problems in measuring change. Madison: The University of Wisconsin Press, 1967.

Pendl, P. Effektivität des sprachlabors an höheren schulen. Unpublished doctoral dissertation, University of Vienna, 1976.

Platzer, H. *Der einfluss technischer spielzeuge auf das mechanischtechnische verständnis*. Unpublished doctoral dissertation, University of Vienna, 1978.

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut, 1960.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, *19*, 49-57.

Rasch, G. An informal report on a theory of objectivity in comparisons. In L. van der Kamp & C. A. J. Vlek (Eds.), *Psychological measurement theory*. Leyden: University of Leyden, Psychological Institute, 1967. (Proceedings of the NUFFIC international summer session in science at Het Oude Hof, The Hague, July 1966.)

Rasch, G. Objectivitet i samfundsvidenskaberne et metodeproblem. *Nationaloekonomisk Tidsskrift*, 1972, *110*, 161-196.

Rella, E. *Trainierbarkeit des antizipierens von gefahrensituationen im strassenverkehr*. Unpublished doctoral dissertation, University of Vienna, 1976.

Rop, I. The application of a linear logistic model describing the effects of pre-school curricula on cognitive growth. In H. Spada & W. F. Kempf (Eds.), *Structural models of thinking and learning*. Bern: Huber, 1977.

Vodopiutz, A. *Komplexbildung im motorischen lernen*. Unpublished doctoral dissertation, University of Vienna, 1977.

Witek, J. *Die Effektivität des gruppendynamischen sensitivity trainings*. Unpublished doctoral disseration, University of Vienna, 1979.

Zeman, M. *Die wirksamkeit der mathematischen früherziehung*. Unpublished doctoral dissertation, University of Vienna, 1976.

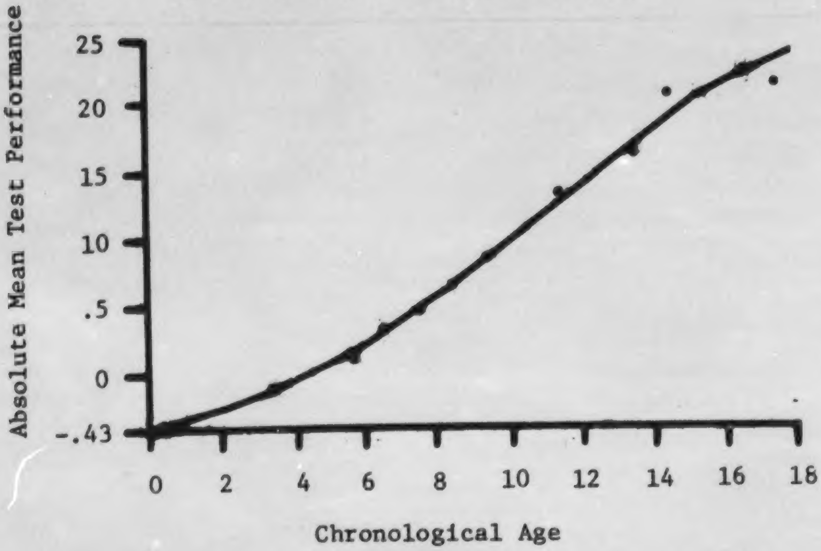
Zimprich, H. *Behandlungskonzepte und resultate bei psychosomatischen erkrankungen im kindesalter*. Unpublished manuscript, 1979.

THE MENTAL GROWTH CURVE RE-EXAMINED

R. DARRELL BOCK
UNIVERSITY OF CHICAGO

A study purporting to show the growth of mental ability, as measured by the Binet test, as a function of chronological age was published in 1929 by Thurstone and Ackerson. The curve was published on a rescaling of Binet mental ages (MA) of a cross-sectional sample of 4,208 children from ages 3 through 17, seen at the Institute for Juvenile Research in Chicago. The shape of that curve, which is reproduced in Figure 1, is surprising in one respect: It shows an inflection point at about 10 years of age, where an initial positive acceleration switches to negative. There is no precedent for this type of growth curve in any other aspect of human growth. All other such curves--in particular, those for growth in stature (see Bock & Thissen, 1980)--show a rapid deceleration from birth through adolescence, followed by a brief period of acceleration during the adolescent growth spurt. (In longitudinal growth records of individual chil-

Figure 1
Thurstone's Curve for Binet Mental Growth
(from Thurstone & Ackerson, 1929)



dren, a slight middle-childhood spurt can sometimes also be seen between 6 and 7 years, but this is not evident in cross-sectional data.)

Any discussion of the shape of such curves requires that the unit of scale be equal at all points throughout the range of measurement. Because there is no reason to suppose that MA scores for the Binet have this property, some method of scaling the test responses that will yield a uniform unit must be adopted. Thurstone (1925,1927,1928) formulated such a method. It rests on two very general assumptions: (1) that the distributions of mental age (or attainment) conditional upon chronological age have the same (continuous) functional form at all age levels but may differ in mean and dispersion (standard deviation); (2) that the origin of measurement can be assigned so that the dispersion of the conditional distributions is directly proportional to the mean, that is, so that the coefficient of variation is constant.

Thurstone pointed out that if the functional form of the common distribution is known, these assumptions may be checked (1) by converting the observed proportions of people at each age level who respond correctly to each test item to the corresponding percentage point of distribution and (2) by plotting the resulting transformed proportions as a function of age. If the points tend to lie on straight lines and the slopes of the lines decrease with increasing age, the assumptions are justified. Thurstone (1925,1927,1928) exhibited numerous examples of data in which these assumptions seem reasonable when the conditional distributions are assumed normal. He also developed simple numerical methods for estimating the item means (thresholds) and the constants of proportionality for the item standard deviations. He called this procedure the "method of absolute scaling." Although the method is no longer used, it is important as a forerunner of modern item characteristic curve (ICC) scaling procedures.

However, this method was not used directly on the item data by Thurstone and Ackerson (1929); rather, they obtained the means and constants of proportionality indirectly from the MA distributions of yearly age groups. (In supplementary tables, the actual data distributions are given in 3-month intervals for boys and girls separately, with boys substantially outnumbering girls in the sample.) This labor-saving compromise of the absolute scaling method can be justified on grounds that the mean and dispersion obtained from the average percent correct for items represented in the MA score will be a good approximation to the average of the means and dispersions of the separate items. There is no reason to believe that the unusual characteristics of the Thurstone-Ackerson curve for mental growth curve are due to their scaling the Binet data at the score level rather than at the item level.

A more plausible explanation is that the shape of the curve is influenced by Thurstone's use of the observed ratios of MA dispersions in successive chronological age groups to determine the factor of proportionality (coefficient of variation) relative to the mean scaled mental age. The growth curve thus obtained, although independent of the arbitrary Binet MA scale in the conditional means, is not independent of the scale in the calculation of the conditional dispersions. A solution independent of arbitrary scale artifacts in both item thresholds and dispersions was not practical with the hand methods of computation then available to Thurstone.

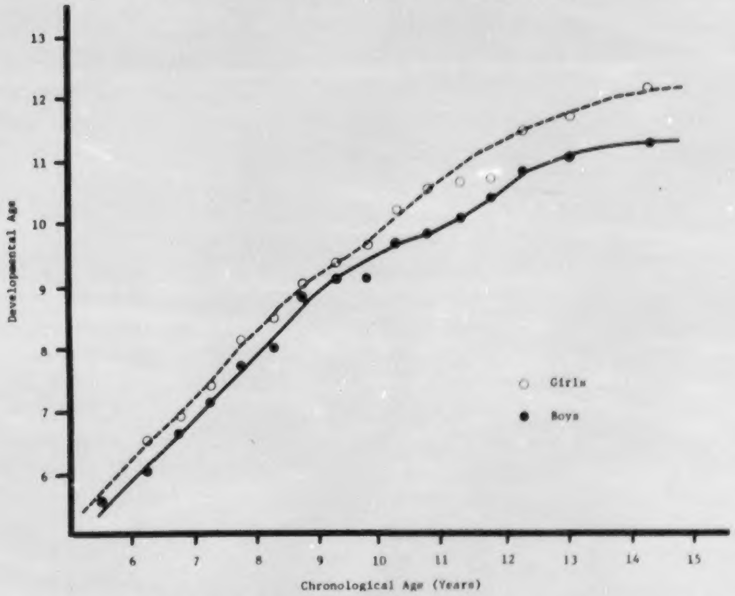
A Scaling Procedure Fully Independent of Chronological Age Units

With the aid of modern computers, however, the Binet test, or similar tests referenced to chronological age or to other external criterion, can be scaled on the single assumption that the underlying distributions for item attainments conditional on age have a known common functional form indexed by a threshold and a dispersion parameter. If this assumption is satisfied, scale values may be assigned to the chronological age groups so that with respect to growth continuum, all the ICCs simultaneously fit the observed percent-correct data for each item in each age group. On the further assumption that item responses within the age groups are independent (locally independent), the goodness of fit of the solution can be tested by a large-sample statistical test.

A Biological Example

A maximum likelihood procedure for scaling by this method, when a normal ogive ICC is assumed, is presented in the appendix to Bock (1976). This procedure has been applied by Kolakowski and Bock (in press) to biological data consisting of counts of emerged permanent dentition in a large cross-sectional sample of Pima Indian children (Dahlberg & Menegaz-Bock, 1958). Reproduced in Figure 2 are the scale values obtained by Kolakowski and Bock (in press), plotted as a function of chronological age. As can be seen, the curves based on incidence of emerged permanent teeth initially decelerate and show some suggestion

Figure 2
Developmental Age Curves Inferred from Emergence of Permanent Dentition in Pima Indian Children (from Kolakowski & Bock, 1980)



of an adolescent growth spurt in both sexes. There is no evidence of the initial positive acceleration that was found in the mental growth curve by Thurstone and Ackerson (1929).

An unavoidable limitation of all such scaling methods is that the origin and unit of measurement of the scale is arbitrary in each sample analyzed. In the case of the tooth emergence data, Kolakowski and Bock (in press) adjusted the origin and unit so that the threshold and dispersion of one of the teeth that is known to show no sex difference in emergence time, an upper central incisor, had the same values as those in the literature based on probit analyses of tooth frequencies as a function of chronological age (Dahlberg & Menegaz-Bock, 1958). The curves for the two sexes in Figure 2 are based on this choice of origin and unit. Thurstone and Ackerson (1929) based the origin of their scale at an inferred point of zero variability (Thurstone's, 1928,"absolute zero" of intelligence) and set the unit so that the MA of the year group equaled chronological age (CA).

Scaling the Binet Test

Data and Method

Using data supplied by Reckase (1979), the Bock (1976) procedure was applied to 96 items of the current version of the Stanford-Binet. These data, which are reproduced in Appendix Tables A and B, are drawn from the full complement of 122 Binet tasks, with the first 13 omitted because all subjects responded correctly and the last 13 omitted because all subjects responded incorrectly.

The numbers and mean age of boys and girls in each CA group are shown in Table 1. In some instances, alternative forms of an item were treated in the scaling as if they were the same item. The data are strictly cross-sectional and, like all such data, are not constrained to be increasing with chronological age (see Bock, 1979).

The scaling solutions based on Bock (1976) converged in 13 Newton-Raphson iterations. The computations required 63 seconds of IBM 370/168 cpu time and 465K bytes of core storage. Scale values for successive 10-month chronological age groups were calculated. The origin and unit of the scale for boys were fixed so that the values for the 40-month and 160-month groups were 40 and 160, respectively. The unit of the girls' scale was then set so that the averages of the item dispersions for boys and girls were equal (to 19.55); and the origin of the girls' scale was set so that the scale value of the 160-month group was 160.

Results: The Revised Mental Growth Curve

The growth curves from this scaling solution are shown in Figure 3. The curve for boys is entirely plausible as a representation of growth. Unlike the Thurstone-Ackerson curve, it decelerates from the earliest age until adolescence. The final two points suggest the possible upward inflection of a slight adolescent growth spurt in mental attainment. At age 14 the curve is still rising, and presumably would go higher if older age groups were included.

Table 1
Mean Chronological Age (CA) and Sample Size
for Each Age Group

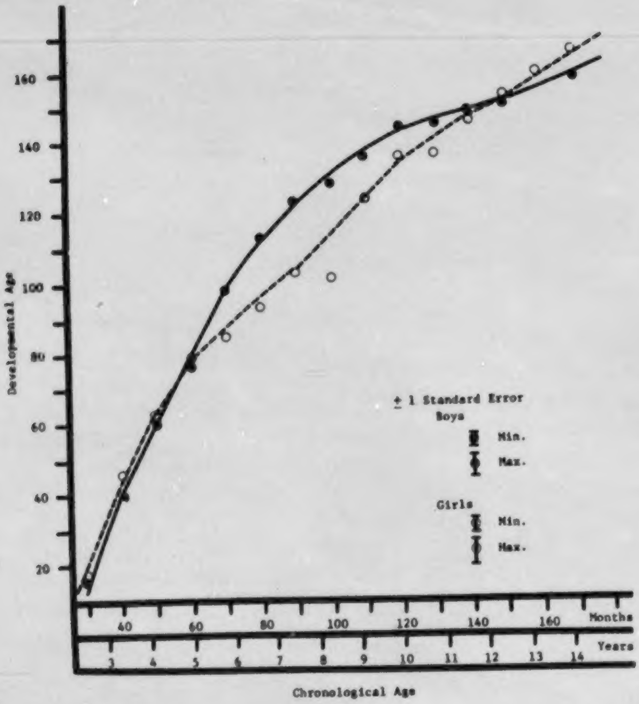
Age Group	CA Interval in Months	Boys (N=342) Mean	Boys N	Girls (N=81) Mean	Girls N
1	24-36	30.9	17	31.8	10
2	37-46	42.6	26	40.7	30
3	47-56	50.5	29	51.4	25
4	57-66	61.3	32	60.9	28
5	67-76	71.4	35	72.2	22
6	77-86	81.7	22	81.4	21
7	87-96	91.5	29	91.6	18
8	97-106	102.1	25	101.3	18
9	107-116	111.6	24	111.5	23
10	117-126	121.3	25	121.7	15
11	127-136	132.1	19	130.6	16
12	137-146	141.3	12	141.6	20
13	147-156	151.3	15	151.3	8
14	157-166	162.0	16	160.5	13
15	167-178	170.9	16	173.1	14

The curve for girls is less satisfactory. Initially, it resembles the curve for boys; but from years 6 through 11, the scale values for girls are irregular and considerably below those for boys of the same age. It is possible, of course, that the equating of boys and girls at 160 months is unfair to the girls. Perhaps they are actually 10 or 20 points higher at that age. If so, the points in the range 70 to 130 months would be more comparable in boys and girls.

Such an adjustment, however, would make the girls' scores in the range 30 to 60 high relative to those of the boys. Inasmuch as the percents correct for girls on items in this range, or indeed in the upper range of 140 to 170 months, were about the same as those for the boys, this interpretation does not seem plausible (compare Tables A and B). The assumption that boys and girls have the same average Binet attainment at 160 months seems reasonable for these data.

The only explanation for the anomalous result for girls would seem to be that the samples for the two sexes were not comparable in some age groups. Some bias in selection of subjects or in administration of the tests must have operated against girls in the 70- to 130-month range. The irregularity of the girls' scale values in this range, especially the discrepant value at 100 months, suggests that the sample of girls may have been defective. Regrettably, no information is available on how the subjects were selected or how the tests were administered.

Figure 3
Proposed Mental Growth Curve Based on Binet
Item Data Collated by Reckase (1979)



Advantages of the Present Scaling Procedure

Because the present scaling method does not force the dispersions of the conditional distributions to increase with age, the scale is not stretched to the left in order to make the conditional standard deviations small. It is this stretching of the scale that induces the initial positive acceleration in the Thurstone-Ackerson curve. When the dispersions were estimated without constraint, the more plausible initial negative acceleration seen in Figure 3 is obtained.

As discussed in Bock (1976) and Kolakowski and Bock (in press), the item parameters estimated in the scaling solutions can also be used to assign developmental age scores to individual subjects by the method of maximum likelihood (see also Birnbaum, 1968; Samejima, 1969). In this role the present scaling solution has important methodological advantages. On the developmental-age scale, the item dispersions, rather than increasing as Thurstone had assumed, are relatively homogenous. A solution with all item standard deviations set to their average value fit almost as well as the unconstrained solution. This implies that the maximum likelihood estimates of developmental age of individual subjects can be expressed with good accuracy as a function of the subject's number-correct score. This is implied by the close similarity of a 1-parameter normal ogive model with the 1-parameter logistic model in which number correct is the sufficient statistic for the maximum likelihood estimate (Andersen, 1980.)

Moreover, when the within-age group standard deviations of the estimated developmental age scores were calculated (Table 2), they were also relatively homogeneous. This means that analysis of variance can be employed to investigate relationships between developmental age and other age-structured data without violating the assumption of homoscedasticity. The conventional MA scores for the Binet do not have this property.

Table 2
Developmental Age Means and
Standard Deviations for Children
in Successive Chronological Age (CA) Groups

Age Group	Nominal CA	Boys		Girls	
		Mean	SD	Mean	SD
1	30	15.4	16.6	16.9	18.1
2	40	40.0	19.2	46.0	12.6
3	50	61.0	18.0	62.4	14.1
4	60	76.2	17.3	76.9	22.1
5	70	96.6	13.5	84.1	10.0
6	80	110.9	10.1	92.1	10.2
7	90	120.6	13.7	101.5	10.1
8	100	125.8	13.2	105.0	11.5
9	110	134.0	13.4	121.1	12.3
10	120	143.3	10.7	134.0	16.7
11	130	144.3	15.6	134.7	16.4
12	140	148.3	9.7	143.9	14.9
13	150	152.1	17.4	153.0	19.9
14	160	160.0	8.2	160.0	19.1
15	170	162.0	16.3	175.1	38.0

The developmental age scale may also have certain interpretational advantages whenever changes within subjects rather than normative comparisons with age-mates is at issue. Because the developmental age units are greater than chronological units at young ages, the changes in scale values are more in accord with the rate of behavioral change (i.e., the surpassing of successive developmental tasks) than with changes in MA scores. Moreover, the growth of mental attainment in terms of scaled scores will parallel closely other quantitative indices of development, such as stature. Thus, the developmental age scores will tend to show simple linear relationships with direct measures of development.

The present scale, however, does not exhibit increasing standard deviation with age and thus does not support Thurstone's definition of the absolute zero of intelligence. Provisionally, at least, it will be necessary to set the origin of the scale on some more arbitrary basis.

REFERENCES

Andersen, E. B. Discrete statistical models with social science applications. Amsterdam: North-Holland, 1980.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading MA: Addison-Wesley, 1968.

Bock, R. D. Basic issues in the measurement of change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), Advances in psychological measurement. London: Wiley & Sons, 1976.

Bock, R.D. Univariate and multivariate analysis of time-structured data. In J. R. Nesselroade & P. B. Baltes (Eds.), Longitudinal research in the study of behavior and development. New York: Academic Press, 1979.

Bock, R. D., & Thissen, D. Statistical problems of fitting individual growth curves. In F. Johnston, A. F. Rocher, & C. Susanne (Eds.), Methodologies for the analysis of human growth and development. New York: Plenum, 1980.

Dahlberg, A. A., & Menegaz-Bock, R. M. Emergence of the permanent teeth in Pima Indian children. Journal of Dental Research, 1958, 37, 1123-1140.

Kolakowski, D., & Bock, R. D. A multivariate generalization of probit analysis. Biometrics, in press.

Reckase, M. Item response data for the Binet tests administered to a large sample of boys and girls age 30 to 170 months. Personal communication, 1979.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, 34 (4, Pt. 2, Monograph No. 17).

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Educational Psychology, 1925, 16, 433-451.

Thurstone, L. L. The unit of measurement in educational scales. Journal of Educational Psychology, 1927, 18, 505-524.

Thurstone, L. L. The absolute zero in intelligence measurement. Psychological Review, 1928, 35, 175-197.

Thurstone, L. L., & Ackerson, L. The mental growth curve for the Binet tests. Journal of Educational Psychology, 1929, 20, 569-583.

ACKNOWLEDGMENT

Preparation of this report was supported in part by NSF Grants BNS-791247 and BNS 76-02849 to the University of Chicago.

Table A
Number of Correct Responses to Each Item for Boys In Successive Age Groups

Item	Age Group 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
14	14	22	29	32	35	22	29	25	24	25	19	12	15	16	15
15	10	21	29	32	35	22	29	25	24	25	19	12	15	16	15
16	12	24	28	32	35	22	29	25	24	25	19	12	15	16	15
17	11	23	27	32	35	22	29	25	24	25	19	12	15	16	15
18	10	25	28	32	35	22	29	25	24	25	19	12	15	16	15
19	10	22	26	29	35	22	29	25	24	25	19	12	15	16	15
20	12	23	27	32	35	22	29	25	24	25	19	12	15	16	15
21	9	21	28	32	35	22	29	25	24	25	19	12	15	16	15
22	8	20	25	30	35	22	29	25	24	25	19	12	15	16	15
23	7	22	26	32	35	22	29	25	24	25	19	12	15	16	15
24	4	19	27	32	35	22	29	25	24	25	19	12	15	16	15
25	3	15	27	30	34	22	29	25	24	25	19	12	15	16	15
26	7	17	25	32	35	22	29	25	24	25	19	12	15	16	15
27	5	13	24	29	35	22	29	25	24	25	19	12	15	16	15
28	6	22	25	30	35	22	29	25	24	25	19	12	15	16	15
29	3	17	24	30	35	22	29	25	24	25	19	12	15	16	15
30	3	12	23	29	35	22	29	25	24	25	19	12	15	16	15
31	4	11	23	24	34	22	29	25	24	25	19	12	15	16	15
32	2	5	22	27	35	22	29	25	24	25	19	12	15	16	15
33	3	16	23	27	34	22	29	25	24	25	19	12	15	16	15
34	1	3	20	27	35	22	29	25	24	25	19	12	15	16	15
35	1	11	18	26	34	22	29	25	24	25	19	12	15	16	15
36	2	12	22	27	35	22	29	25	24	25	19	12	15	16	15
37	0	8	12	21	31	22	29	25	24	25	19	12	15	16	15
38	0	3	13	23	32	21	29	25	24	25	19	12	15	16	15
39	1	11	22	27	34	22	29	25	24	25	19	12	15	16	15
40	0	3	16	22	34	22	28	25	24	25	19	12	15	16	15
41	1	5	20	22	32	22	29	25	24	25	19	12	15	16	15
42	0	4	7	17	24	22	28	25	23	25	19	12	15	16	15
43	0	1	7	16	29	21	27	24	24	25	19	12	15	16	15
44	0	2	5	12	31	20	28	25	24	25	19	12	15	16	15
45	0	3	5	22	29	22	29	25	23	25	19	12	15	16	15
46	0	0	4	13	28	22	29	25	24	25	19	12	15	16	15
47	0	0	7	22	24	22	27	25	24	25	19	12	15	16	15
48	0	1	8	18	31	22	29	25	23	25	19	12	15	16	15
49	0	0	0	9	18	17	24	22	24	25	19	12	15	16	15
50	0	0	0	4	19	17	26	24	24	25	19	12	15	16	15
51	0	0	1	5	21	20	27	24	23	25	19	12	15	16	15
52	0	0	0	7	23	17	27	25	22	24	19	12	15	16	15
53	0	0	4	8	21	17	26	23	24	24	19	12	15	16	15
54	0	0	1	2	11	17	25	20	23	25	17	12	15	16	15
55	0	0	0	5	19	16	26	23	23	25	19	12	15	16	15
56	0	0	0	5	12	16	24	22	20	24	19	12	15	16	15
57	0	0	0	0	7	13	21	17	21	24	18	11	14	16	15
58	0	0	0	1	13	13	24	21	23	25	18	11	15	16	15
59	0	0	0	1	13	10	23	22	19	24	18	11	14	16	15
60	0	0	0	2	7	15	23	22	22	24	19	12	14	16	15
61	0	0	0	1	9	6	17	17	21	23	18	12	15	16	15
62	0	0	0	0	2	6	18	18	21	21	17	11	13	16	15
63	0	0	0	1	5	9	15	15	19	25	18	12	14	16	15
64	0	0	0	0	6	13	21	20	23	24	19	12	14	16	15
65	0	0	0	0	0	3	20	15	21	25	18	12	14	16	15
66	0	0	0	1	4	7	17	18	18	23	16	10	13	16	15
67	0	0	0	0	2	2	9	15	18	22	19	12	14	16	15
68	0	0	0	0	2	4	11	11	17	22	16	11	13	16	15
69	0	0	0	1	1	2	12	13	16	19	16	12	14	16	15
70	0	0	0	0	1	5	14	13	16	20	15	12	13	16	15
71	0	0	0	0	4	7	15	15	19	21	17	11	13	16	15
72	0	0	0	0	2	5	13	11	17	21	14	10	14	16	14
73	0	0	0	1	1	3	5	10	10	18	13	11	12	15	15
74	0	0	0	0	0	2	4	6	14	19	12	10	12	16	14
75	0	0	0	0	0	2	8	9	17	19	14	12	13	16	15
76	0	0	0	0	1	4	11	12	12	18	12	8	13	15	13
77	0	0	0	0	1	0	8	11	11	17	13	11	13	16	15
78	0	0	0	0	1	3	10	12	20	23	14	11	15	16	14
79	0	0	0	0	0	0	3	5	7	16	12	10	11	15	14
80	0	0	0	0	0	1	9	6	13	17	12	10	11	14	15
81	0	0	0	0	0	1	5	3	11	18	10	8	11	16	13
82	0	0	0	0	0	0	2	8	12	14	12	8	12	14	13
83	0	0	0	0	0	0	4	6	6	15	11	12	12	16	14
84	0	0	0	0	0	0	2	5	10	16	12	7	11	13	14
85	0	0	0	0	0	0	6	5	11	14	11	5	9	9	11
86	0	0	0	0	0	0	1	8	9	15	11	10	12	15	13
87	0	0	0	0	0	0	3	5	6	11	10	6	10	11	12
88	0	0	0	0	0	1	5	3	11	15	12	7	10	14	14
89	0	0	0	0	0	0	0	5	6	14	9	9	10	12	12
90	0	0	0	0	0	0	1	6	7	14	11	6	11	14	14
91	0	0	0	0	0	0	2	1	5	11	8	8	11	13	11
92	0	0	0	0	0	0	1	2	4	10	8	4	8	12	12
93	0	0	0	0	0	0	1	1	8	9	7	6	11	12	10
94	0	0	0	0	0	0	1	1	7	8	5	7	8	15	11
95	0	0	0	0	0	1	1	4	7	13	11	9	11	15	14
96	0	0	0	0	0	1	2	6	8	17	9	7	13	12	14
97	0	0	0	0	0	0	0	0	0	3	4	3	4	10	6
98	0	0	0	0	0	0	0	0	1	0	4	2	7	10	7
99	0	0	0	0	0	0	0	0	0	2	4	2	6	13	7
100	0	0	0	0	0	0	0	1	2	6	8	2	7	11	9
101	0	0	0	0	0	0	0	1	1	6	4	1	5	10	9
102	0	0	0	0	0	1	1	2	5.	12	8	5	9	14	12
103	0	0	0	0	0	0	0	1	2	6	8	4	4	12	7
104	0	0	0	0	0	0	0	0	0	4	3	5	7	6	7
105	0	0	0	0	0	0	0	0	0	1	2	1	4	4	4
106	0	0	0	0	0	0	0	0	3	3	5	5	1	5	5
107	0	0	0	0	0	0	0	0	1	2	1	1	2	6	5
108	0	0	0	0	0	0	0	0	0	4	6	4	5	7	7
109	0	0	0	0	0	0	0	0	0	3	2	4	5	7	5
Attempts	17	26	29	32	35	22	29	25	24	25	19	12	15	16	15

Table B

Number of Correct Responses to Each Item for Girls in Successive Age Groups

	Age Group														
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
14	9	28	25	28	22	21	18	18	23	15	16	20	8	13	12
15	8	29	24	28	22	21	18	18	23	15	16	20	8	13	12
16	8	29	24	28	22	21	18	18	23	15	16	20	8	13	12
17	6	26	24	28	22	21	18	18	23	15	16	20	8	13	12
18	7	28	24	28	22	21	18	18	23	15	16	20	8	13	12
19	5	26	24	27	22	21	18	18	23	15	16	20	8	13	12
20	4	25	24	28	22	21	18	18	23	15	16	20	8	13	12
21	4	27	24	27	22	21	18	18	23	15	16	20	8	13	12
22	3	28	23	28	22	21	18	18	23	15	16	20	8	13	12
23	7	26	25	28	22	21	18	18	23	15	16	20	8	13	12
24	2	24	24	27	22	21	18	18	23	15	16	20	8	13	12
25	2	21	23	26	22	21	18	18	23	15	16	20	8	13	12
26	3	18	23	26	22	21	18	18	23	15	16	20	8	13	12
27	2	20	22	26	22	21	18	18	23	15	16	20	8	13	12
28	3	26	24	26	22	21	18	18	23	15	16	20	8	13	12
29	3	23	22	27	22	21	18	18	23	15	16	20	8	13	12
30	0	19	20	27	22	21	18	18	23	15	16	20	8	13	12
31	3	19	21	25	22	21	18	18	23	15	16	20	8	13	12
32	1	13	18	24	21	21	18	18	23	15	16	20	8	13	12
33	2	20	23	27	22	21	18	18	23	15	16	20	8	13	12
34	0	12	18	24	22	21	18	18	23	15	16	20	8	13	12
35	3	19	20	24	20	21	18	18	23	15	16	20	8	13	12
36	0	17	19	27	22	21	18	18	23	15	16	20	8	13	12
37	0	7	19	25	22	21	18	18	23	15	16	20	8	13	12
38	1	7	17	26	22	20	18	18	23	15	16	20	8	13	12
39	1	13	20	26	22	21	18	18	23	15	16	20	8	13	12
40	0	2	18	23	22	21	18	18	23	15	16	20	8	13	12
41	0	8	22	27	22	20	18	18	23	15	16	20	8	13	12
42	0	4	10	19	17	20	17	18	23	15	16	20	8	13	12
43	0	6	9	22	21	19	18	18	23	15	16	20	8	13	12
44	0	2	11	20	21	21	18	18	23	15	16	20	8	13	12
45	0	2	11	21	18	21	18	18	23	15	16	20	8	13	12
46	0	1	8	17	19	21	18	18	23	15	16	20	8	13	12
47	0	3	8	20	18	20	18	18	23	15	16	20	8	13	12
48	0	1	12	23	20	21	18	17	23	15	16	20	8	13	12
49	0	0	2	10	9	11	15	16	23	15	16	20	8	13	12
50	0	0	3	8	10	13	17	16	23	15	16	20	8	13	12
51	0	0	0	7	13	19	16	18	23	15	16	20	8	13	12
52	0	0	3	11	13	19	14	17	23	15	16	20	8	13	12
53	0	1	1	14	14	15	18	16	23	15	16	20	8	13	12
54	0	0	3	5	8	13	15	16	23	15	16	20	8	13	12
55	0	0	2	10	13	12	14	17	23	15	16	20	8	13	12
56	0	0	2	7	8	12	16	17	23	15	16	20	8	13	12
57	0	0	0	4	6	7	9	13	23	14	15	20	8	13	12
58	0	0	0	7	8	14	14	16	23	15	16	20	8	13	12
59	0	0	0	8	8	10	12	16	23	15	15	20	8	13	12
60	0	0	0	3	9	10	13	17	23	15	15	20	8	13	12
61	0	0	0	1	1	7	9	10	18	13	16	19	8	13	12
Attempts	10	30	25	28	22	21	18	18	23	15	16	20	8	13	12

	Age Group														
Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
62	0	0	0	1	2	2	8	12	22	13	15	19	8	13	12
63	0	0	0	1	4	11	11	9	17	14	14	19	8	13	12
64	0	0	0	2	5	8	14	14	22	15	16	20	8	13	12
65	0	0	0	1	3	3	7	11	23	15	16	19	8	13	12
66	0	0	0	1	1	9	10	11	18	15	15	20	8	13	12
67	0	0	0	1	3	2	3	12	19	14	16	19	7	13	12
68	0	0	0	1	1	5	9	11	11	12	14	19	7	12	12
69	0	0	0	1	0	3	4	5	18	13	14	20	8	12	12
70	0	0	0	2	1	2	8	10	15	14	13	19	8	12	12
71	0	0	0	2	3	8	11	8	19	14	13	18	8	13	12
72	0	0	0	2	2	7	11	3	15	11	13	19	7	12	12
73	0	0	0	1	0	7	5	5	13	10	11	18	6	13	12
74	0	0	0	1	1	1	4	4	13	10	12	18	7	13	11
75	0	0	0	1	0	2	1	3	16	14	13	17	8	13	12
76	0	0	0	1	2	4	7	2	14	11	12	18	8	11	12
77	0	0	0	1	3	3	4	5	16	8	12	15	8	13	12
78	0	0	0	1	1	2	7	5	17	13	15	19	8	13	12
79	0	0	0	1	0	0	0	2	7	11	11	16	7	12	11
80	0	0	0	1	1	1	5	4	15	12	11	15	7	11	11
81	0	0	0	1	0	2	2	3	2	8	10	14	5	11	12
82	0	0	0	1	0	0	6	4	14	11	12	16	4	13	11
83	0	0	0	1	0	0	1	3	9	11	10	16	7	13	12
84	0	0	0	1	0	2	2	5	13	9	14	16	8	11	12
85	0	0	0	1	0	1	4	1	12	8	7	11	6	12	11
86	0	0	0	1	0	0	0	1	8	10	11	17	7	11	11
87	0	0	0	1	0	1	1	1	7	7	10	11	5	11	11
88	0	0	0	1	0	0	2	1	8	10	13	16	7	12	11
89	0	0	0	1	0	0	2	1	10	10	9	15	7	12	10
90	0	0	0	1	0	1	3	3	10	8	10	15	4	13	11
91	0	0	0	1	0	0	0	0	4	5	7	14	7	12	10
92	0	0	0	1	0	1	0	0	2	7	6	13	3	9	9
93	0	0	0	1	0	0	0	1	4	7	7	13	4	9	11
94	0	0	0	1	0	0	0	0	2	6	4	12	5	9	9
95	0	0	0	1	0	1	0	0	2	6	5	14	6	10	8
96	0	0	0	1	0	1	0	2	6	7	11	18	7	12	10
97	0	0	0	1	0	0	0	0	2	3	2	6	4	9	10
98	0	0	0	1	0	0	0	0	0	2	2	6	3	8	6
99	0	0	0	1	0	0	0	0	2	2	4	5	5	8	10
100	0	0	0	0	0	0	0	0	2	5	4	8	7	9	10
101	0	0	0	1	0	0	0	0	0	2	3	6	6	8	8
102	0	0	0	1	0	0	0	0	3	5	3	11	3	7	6
103	0	0	0	1	0	0	0	0	1	4	4	9	5	9	7
104	0	0	0	1	0	0	0	0	1	3	2	8	5	8	7
105	0	0	0	1	0	0	0	0	0	0	0	2	4	3	2
106	0	0	0	0	0	0	0	0	0	1	4	4	2	3	3
107	0	0	0	0	0	0	0	0	0	1	1	4	2	4	4
108	0	0	0	0	0	0	0	0	1	4	5	6	4	7	6
109	0	0	0	1	0	0	0	0	0	2	2	6	3	8	5
Attempts	10	30	25	28	22	21	18	18	23	15	16	20	8	13	12

LATENT STRUCTURE ESTIMATION FOR ASSESSING GAIN IN ABILITY

LALITHA SANATHANAN
ARGONNE NATIONAL LABORATORY

This paper deals with methods for assessing the progress of an individual or group through time. The methods involve (1) measuring the gain in ability over a given period of time using a latent ability model, such as the Rasch model and (2) relating this gain to the average gain for similar individuals or groups over the same length of time. The changes in ability parameters for individuals and for groups can be estimated through existing methods based on latent trait models. However, in order to judge whether a specific individual or group has progressed satisfactorily, it is necessary to compare the given gain in ability with gains for similar individuals or groups.

It is common practice to report test scores based on a hierarchical test system such as the Iowa Tests of Basic Skills (ITBS) in the form of grade equivalent scores. The grade equivalent of any given test score is approximately the grade whose mean is the given score. Its principal use is to measure the progress of an individual or group over a given period of time. The increase in grade equivalent scores, referred to as the gain score, is considered a measure of this type of longitudinal progress. In spite of numerous problems in the interpretation of grade equivalent scores, the gain score has a certain appeal in that it tries to express progress in terms of gain in years. This paper provides a measure of longitudinal progress that is interpretable in terms of gain in years but overcomes the objections to the use of grade equivalent scores. The measure proposed here is obtained by first using the Rasch model of latent ability to measure gain in ability on a non-normative scale, and then providing a normative interpretation for this gain.

Let θ_1 and θ_2 be the values of the ability parameter for an individual at times $\underline{t}_1$ and $\underline{t}_2$. Given an initial ability level of θ_{10} and a gain of $\theta_{20} - \theta_{10}$ over the period $\underline{t}_2 - \underline{t}_1$, assessment of this gain can be made on the basis of the conditional distribution of θ_2, given $\theta_1 = \theta_{10}$ for a norm group, such as a national sample. In particular, the mean and standard deviation of this conditional distribution enable the expression of an absolute gain as a percentile gain, which in turn has the usual interpretation.

This paper provides an empirical Bayes procedure for computing the parameters of the above conditional distribution needed for this type of judgment. On the basis of the estimated parameters, the time it takes, on the average, for an individual with initial ability θ_{10} to achieve a gain of $\theta_{20} - \theta_{10}$ can also be computed, thus making possible the expression of progress on a chronological scale. Two other related applications of the empirical Bayes procedure are also discussed.

The Rasch Model

The Rasch model can be described as follows: Let θ denote a real-valued parameter representing the ability of an individual, and let $\underline{p}(\theta)$ be the probability that an individual with parameter θ will correctly solve item $\underline{j}$ from a given pool of items. The Rasch model specifies that

$$p_j(\theta) = \exp\{\theta + \phi_j\} \;/\; (1 + \exp\{\theta + \phi_j\}), \quad j = 1,\ldots,m \qquad [1]$$

or, equivalently,

logit $\underline{p}_j(\theta) = \theta + \phi_j$, $j = 1,\ldots,\underline{m}$, where ϕ_j is a real-valued parameter characterizing the difficulty of the $\underline{j}^{th}$ item and $\underline{m}$ is the number of items.

Consider a group of individuals with ability parameters θ_i whose responses to $\underline{j}$ items are observed. Under the assumption that individuals respond independently of one another and that for the same individual, responses to different items are mutually independent, maximum likelihood or other estimates of the θ_i's and ϕ_j's can be obtained (for details see Anderson, 1970; Wright & Panchapakesan, 1969). Assume that the raw scores for an individual at two points in time--$\underline{t}_1$ and $\underline{t}_2$--are based on two different tests, such as those corresponding to different hierarchical levels of a test system. Assuming that the items on the two tests are calibrated and that estimates of the item parameters are available, the raw scores χ_1 and χ_2 would be used separately on the two tests to estimate the abilities θ_1 and θ_2 for the individual at $\underline{t}_1$ and $\underline{t}_2$, respectively. There would thus be an estimate of the gain in ability $\theta_2 - \theta_1$ for the individual over the period $\underline{t}_1$ to $\underline{t}_2$. This measure, however, has very little meaning, unless it is given a normative interpretation. It does not, for instance, denote whether a specific individual has progressed satisfactorily.

In order to make a judgment of this nature, it is necessary to compare $\theta_2 - \theta_1$ for the given individual with gains for similar individuals. The conditional distribution of θ_2, given θ_1 for a norm group provides a useful basis for the above comparison. The mean and standard deviation of this conditional distribution are relevant measures by which the gain $\theta_2 - \theta_1$ in an individual's ability can be judged. This type of comparison involving conditional averages is more appropriate than the one based on grade equivalent scores, since the former takes into account the fact that gain in ability itself is dependent on initial ability level, whereas for the latter, comparison averaging is done over all individuals in certain norm groups without regard to their abilities. For this reason gain expressed in terms of grade equivalent scores is likely to be inflated if an individual with a high initial ability is being considered, the opposite being true in the case of low-ability individuals. Such distortions are avoided by the proposed method based on conditional averages.

An Empirical Bayes Model for Assessing Gain in Ability

The need for estimating the mean and standard deviation of the conditional distribution of θ_2 (ability at time t_2), given θ_1 (ability at time t_1), for a specified group has been established in the previous section. In this section a suitable model and a method for obtaining these estimates is outlined for the estimation process.

The Model

Consider, for instance, a group of individuals whose raw scores, based on different levels of a test, are available at two different times, t_1 and t_2. Let the raw scores for individual i at times t_1 and t_2 be denoted by r_{i1} and r_{i2}, respectively. It is assumed that at each time point the raw scores are adequately described by a Rasch model and that estimates of all the item parameters are available. For the present purpose the item parameters will be treated as if they are known. The tests are not required to be the same for all individuals or to be the same at times t_1 and t_2 for the same individual.

Each individual i can be characterized by θ_{i1}, r_{i1}, S_{i1} and θ_{i2}, r_{i2}, S_{i2}, where θ_{i1} and θ_{i2} are the individual's latent abilities at times t_1 and t_2, respectively; S_1 and S_2 are the sets of item parameters relevant to the two tests taken by the individual; and r_{i1} and r_{i2} are the raw scores defined earlier. It can be further assumed that the sample under consideration is drawn from a population of individuals whose abilities at times t_1 and t_2 follow a bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation coefficient ρ. This type of longitudinal model has been used by Andersen (1979) in another context.

Representing the latent abilities in this population at times t_1 and t_2 by the generic variables θ_1 and θ_2, the joint distribution of θ_1 and θ_2 may then be specified to be bivariate normal with density denoted by $\phi(\theta_1, \theta_2)$. The density $\phi(\theta_1, \theta_2)$ resembles a Bayesian prior density. However, an empirical Bayes approach is followed here in that the parameters of the prior density are estimated from the sample. The conditional distribution of θ_2, given θ_1, is univariate normal with mean and variance

$$E(\theta_2 | \theta_1) = \mu_2 + \frac{\rho(\theta_1 - \mu_1)\,\sigma_2}{\sigma_1}$$

$$\mathrm{Var}(\theta_2 | \theta_1) = {\sigma_2}^2 (1 - \rho^2) \qquad [2]$$

In order to estimate $E(\theta_2|\theta_1)$ and $\mathrm{Var}(\theta_2|\theta_1)$, it is thus sufficient to estimate the parameters μ_1, μ_2, σ_1, σ_2, and ρ of the bivariate density $\phi(\theta_1, \theta_2)$. In this problem θ_1 and θ_2 are latent, or unobservable, variables whose characteristics--namely, μ_1, μ_2, σ_1, σ_2, and ρ--are to be estimated. The estimation of this latent structure must be done on the basis of indirect observations represented by the responses of the individuals in the given sample to items on different tests at t_1 and t_2. A method for estimating latent structure in a similar situation involving a univariate latent ability distribution has been provided by Sanathanan and Blumenthal (1978). An extension of this method, which is discussed in the following section, gives the required estimates for the problem considered here.

Once estimates of μ_1, μ_2, σ_1, σ_2, and ρ are obtained, a specific individual's progress over the period from t_1 to t_2 can be judged as follows: Let θ_{10} and θ_{20} be the θ_1 and θ_2 values, respectively, for a given individual. Compute $E(\theta_2|\theta_{10})$ and $Var(\theta_2|\theta_{10})$, use them to express the absolute gain in ability for the individual as a percentile gain, which in turn has the usual interpretation.

The gain $\theta_{20} - \theta_{10}$ can also be interpreted in terms of gain in years as follows: Given an individual with ability θ_{10}, the expected gain for this individual over the period t_1 to t_2 is $E(\theta_2|\theta_{10}) - \theta_{10}$. Let θ_3 be the ability of an individual at time t_3 where $t_3 - t_2 = t_2 - t_1$. The expected gain for an individual with ability θ_{10} over the period t_1 to t_3 can be computed as follows:

$$\begin{aligned} E(\theta_3|\theta_{10}) &= E_{\theta_2}[E(\theta_3|\theta_{10}, \theta_2)] \\ &= E_{\theta_2|\theta_{10}}[E(\theta_3|\theta_2)] \\ &= \mu_2 + \rho \frac{[E(\theta_2|\theta_{10}) - \mu_1]\,\sigma_2}{\sigma_1} \\ &= E[\theta_2|\theta_1 = E(\theta_2|\theta_{10})] \qquad [3] \end{aligned}$$

Thus, for an individual with initial ability θ_1, $E(\theta_2|\theta_1)$--and hence expected gain for the period $t_2 - t_1$ or any multiple thereof--can be computed. The expected gains can then be plotted against the corresponding time periods. Given that an individual with ability θ_{10} has gained $\theta_{20} - \theta_{10}$ in ability, initial ability θ_{10} can be looked up in the expected gain chart and the time period corresponding to an expected gain of $\theta_{20} - \theta_{10}$ can be determined by interpolation. This time period can be interpreted as the gain in time for the individual. Depending on whether this gain is less or greater than $t_2 - t_1$, the individual can be considered as below or above average in performance.

Latent Structure Estimation

This section focuses on the estimation of the parameters μ_1, μ_2, σ_1, σ_2, and ρ of the bivariate density $\phi(\theta_1, \theta_2)$. As pointed out above, these estimates provide the necessary information for assessing the gain $\theta_{20} - \theta_{10}$ in an individual's ability. (This gain itself is estimated on the basis of the individual's responses to items on two different tests and an assumption of the Rasch model.)

Let the test responses for an individual be represented by vectors V_1 and V_2 corresponding to the tests at t_1 and t_2, respectively. Let the j^{th} component of $\underline{V}_k$ be 1 if the $\underline{j}^{th}$ item on the $\underline{k}^{th}$ test is solved correctly, and 0 otherwise. The response vectors can be thought of as being generated by a sequence of independent, identically distributed latent random vectors $(\theta_{i1}, \theta_{i2})$, each with a bivariate normal distribution with density $\phi(\theta_1, \theta_2)$. Since estimating the parameters of $\phi(\theta_1, \theta_2)$ is of interest, the estimation would ideally be based on the pairs $(\theta_{i1}, \theta_{i2})$. In that case, the maximum likelihood method would yield the following estimates:

$$\hat{\mu}_1 = \frac{\Sigma\, \theta_{i1}}{n}$$

$$\hat{\mu}_2 = \frac{\Sigma\, \theta_{i2}}{n}$$

$$\hat{\sigma}_1^2 + \hat{\mu}_1^2 = \frac{\Sigma\, \theta_{i1}^2}{n}$$

$$\hat{\sigma}_2^2 + \hat{\mu}_2^2 = \frac{\Sigma\, \theta_{i2}^2}{n}$$

$$\hat{\rho} = \frac{\Sigma\, \theta_{i1}\, \theta_{i2}}{\hat{\sigma}_1\, \hat{\sigma}_2} \qquad [4]$$

However, since the $(\theta_{i1}, \theta_{i2})$'s are not directly observable, the indirect observations must be relied upon, namely, the response vectors to make the appropriate inferences; and it is plausible to substitute $E(\theta_{ik}|V_{1i}, V_{2i})$ for θ_{ik}, $E(\theta_{ik}^2|V_{1i}, V_{2i})$ for θ_{ik}^2, and $E(\theta_{i1}\, \theta_{i2}|V_{1i}, V_{2i})$ for $\theta_{i1}\, \theta_{i2}$ in Equation 4. This is the approach followed here in estimating μ_1, μ_2, σ_1, σ_2, and ρ. The approach is based on the missing information principle (MIP) formulated by Orchard and Woodbury (1972) and yields the maximum likelihood estimates of the parameters in question. The rationale for the MIP approach is provided here in an intuitive sense. A rigorous explanation is provided by Sanathanan and Blumenthal (1978), on the basis of which it is evident that an application of MIP in this situation does lead to maximum likelihood estimates.

The conditional expectations, such as $E(\theta_{ik}|V_{1i}, V_{2i})$, which are to be substituted for the corresponding latent variables in Equation 4 depend on the values of the parameters μ_1, μ_2, σ_1, σ_2, and ρ, which are themselves unknown and are to be estimated. The MIP approach requires that the values of these parameters and those satisfying Equation 4 be the same. This equality can be achieved through the following iterative procedure, referred to as the EM algorithm by Dempster, Laird, and Rubin (1977), who also show the convergence of this type of algorithm in a much more general setting.

Starting with trial values for μ_1, μ_2, σ_1, σ_2, and ρ, cycle through the E- and M-steps given below, until convergence is attained.

E-step: Compute the conditional expectations such as $E(\theta_{ik} | V_{i1}, V_{2i})$, using the current values of the parameters μ_1, μ_2, σ_1, σ_2, and ρ.

M-step: Revise the parameter values by using Equation 4 and the conditional expectations from the E-step in place of the latent variables appearing in Equation 4.

Let $g_i(\theta_1, \theta_2)$ be the density of $(\theta_{i1}, \theta_{i2})$, conditional on the response vector (V_{i1}, V_{2i}). Then $g_i(\theta_1, \theta_2)$ is given by

$$g_i(\theta_1, \theta_2) = \frac{\phi(\theta_1,\theta_2) \prod_{k=1}^{2} \prod_{jk=1}^{m_k} (p_{jk}(\theta_k))^{x_{ijk}} (1 - p_{jk}(\theta_k))^{1 - x_{ijk}}}{\int_{-\infty}^{\infty} \phi(\theta_1,\theta_2) \prod_{k=1}^{2} \prod_{jk=1}^{m_k} (p_{jk}(\theta_k))^{x_{ijk}} (1 - p_{jk}(\theta_k))^{1 - x_{ijk}} d\theta_1 \, d\theta_2} \qquad [5]$$

where

$\phi(\theta_1,\theta_2)$ is the bivariate normal density,
k is the test number,
m_k is the number of items on the k^{th} test,
$P_{jk}(\theta)$ is given by Equation 1, and
x_{ijk} is 1 if the j^{th} item on the k^{th} test is answered correctly by the i^{th} individual, and is 0 otherwise.

In addition,

$$E(\theta_{ik} | V_{i1}, V_{i2}) = \int \theta_k \, g_i(\theta_1, \theta_2) \, d\theta_1 \, d\theta_2 \qquad , k = 1,2$$

$$E(\theta^2_{ik} | V_{i1}, V_{i2}) = \int \theta_k^2 \, g_i(\theta_1, \theta_2) \, d\theta_1 \, d\theta_2 \qquad , k = 1,2$$

and

$$E(\theta_{i1} \theta_{i2} | V_{i1}, V_{i2}) = \int \theta_{i1} \theta_{i2} \, g_i(\theta_1, \theta_2) \, d\theta_1 \, d\theta_2 \quad . \qquad [6]$$

If each of the two tests at times $\underline{t}_1$ and $\underline{t}_2$ are the same for all individuals, then $\underline{g}_i(\theta_1, \theta_2)$ is the same for all individuals with the same raw scores $(\underline{r}_{i1}, \underline{r}_{i2})$. It is then enough to consider the conditional expectations in Equation 6 for all possible pairs of raw scores. On the other hand, if the individuals are administered different tests at any particular time, then the conditional expectations in Equation 6 must be evaluated separately for each individual. Here again the individual's raw scores on the two tests are sufficient for evaluating the conditional expectations in Equation 6. Basically, the expressions in Equation 6 can be rewritten by noting that

$$g_i(\theta_i, \theta_2) = \frac{\phi(\theta_1, \theta_2) \exp \{\theta_1 r_{i1} + \theta_2 r_{i2}\}}{\prod_{k=1}^{2} \prod_{jk=1}^{m_k} (1 + \exp \{\theta_k + \phi_{jk}\})} , \qquad [7]$$

where ϕ_{jk} is the item parameter for the $\underline{j}^{th}$ item on the $\underline{k}^{th}$ test taken by the $\underline{i}^{th}$ individual and is assumed to be known (or estimated separately).

Computing the expectations in Equation 6 calls for numerical integration, which is done by using the FORTRAN version of CACM Algorithm 145 called ASIMPS. This is the same program that was used for the computations described by Sanathanan and Blumenthal (1978).

A remark concerning the accuracy of estimation is in order. As in regression analysis, for the estimation of $E(\theta_2|\theta_{10})$ the best accuracy is obtained when θ_{10} is the same as or close to the mean ability μ_1 of the group used for estimating the parameters of $\phi(\theta_1, \theta_2)$. For adequate estimation, there must therefore be several samples of which the mean abilities are spread over the range of interest. For a given initial ability θ_{10}, $E(\theta_2|\theta_{10})$ would then be computed using estimates of parameters based on the sample whose mean ability is closest to θ_{10}.

Numerical Illustration

The procedure which has been described for estimating the parameters μ_1, μ_2, σ_1, σ_2, and ρ, of the bivariate density $\phi(\theta_1, \theta_2)$ is illustrated using the following synthetic data. Table 1 represents the responses of 1,000 individuals to tests at two different times. Each test consists of four items whose ϕ_j estimates are given in Table 2.

The maximum likelihood solution is obtained by an iterative procedure involving all five parameters simultaneously. A computational shortcut is achieved here by obtaining estimates for μ_1 and σ_1, and μ_2 and σ_2 separately, based on the respective marginal distributions. Although this procedure is not strictly valid by the maximum likelihood criterion, it is an acceptable compromise between computational efficiency and theoretical rigor. The computational procedures used in obtaining these estimates are as follows: Trial values for μ_1, σ_1 and μ_2, σ_2 were chosen as .1 and 1.0, respectively. In each case, the initial values for μ and σ and the relevant ϕ_j values, and the relevant marginal

Table 1
Bivariate Frequency Distribution of Raw Scores

Time t_1 Raw Score	Time t_2 Raw Score 0	1	2	3	4	Row Margin
0	32	25	22	18	2	99
1	53	69	42	40	8	212
2	33	75	104	95	7	314
3	25	29	139	146	9	348
4	1	3	7	5	11	27
Column Margin	144	201	314	304	37	

raw score frequencies were entered into a computer program for carrying out the E- and M-steps outlined above. This part of the computation involves only the respective conditional means and variances (and not covariances) and marginal distributions of θ_1 and θ_2 separately. After five iterations the final estimates of μ_1 and σ_1 obtained were -.32 and .839, respectively. The estimates for μ_2 and σ_2 were obtained after two iterations as .14 and 1.007, respectively.

Table 2
ϕ_j Estimates of Test Items

Item	Test 1	Test 2
1	-.4033	-1.5921
2	.4476	.3064
3	.4791	-1.0051
4	.6743	1.0932

For estimating ρ the values of μ_1, σ_1 and μ_2, σ_2 were treated as if known, and their estimates were inserted into the expression for $\underline{g}(\theta_1, \theta_2)$, the generic density of (θ_1, θ_2) conditional on a given pair of raw scores. A trial value of $\rho = .8$ was used for evaluating the expectation of (θ_1, θ_2) conditional on various combinations of raw scores, constituting the E-step. The average of these conditional expectations was, in turn, used to revise the value of ρ, as required by the M-step. After two iterations, the ρ estimate obtained was .6.

Related Applications

The empirical Bayes procedure used in assessing longitudinal progress can also be applied to the following related problems:

Consider the problem of evaluating the effectiveness of a new program or a new instructional method. There are usually an experimental group and a control group that are to be compared on the basis of "before" and "after" test scores.

Since gain in ability is, to some extent, dependent on initial ability level, for a meaningful comparison differences in the initial ability levels of the groups must be considered. This can be accomplished as follows: Estimate $E(\theta_2|\theta_1)$ for the groups separately and average the resulting functions over θ_1, using a common marginal distribution for θ_1 (this could, for instance, be the ability distribution of some specified norm group). The averages thus obtained would be free of biases resulting from differences in initial ability levels and hence are comparable.

Another problem to which the empirical Bayes procedure presented in this paper is applicable is that of estimating the correlation coefficient between two tests intended to measure the same or possibly different latent traits. To do this, let the latent traits to be measured by the two tests correspond to θ_1 and θ_2 in the empirical Bayes model and follow the procedure described for computing the required correlation coefficient. This approach circumvents the difficulties encountered in the usual approach, where θ_1 and θ_2 are first estimated for each individual in a given sample and the resulting estimates are used for computing the correlation coefficient.

REFERENCES

Andersen, E. B. Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society, Series B, 1970, 32, 283-301.

Andersen, E. B. Comparing latent distributions. Manuscript submitted for publication, 1979.

Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 1977, 39, 1-38.

Orchard, T., & Woodbury, M. A. A missing information principle: Theory and applications. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, I. Berkeley: University of California Press, 1972.

Sanathanan, L., & Blumenthal, S. The logistic model and estimation of latent structure. Journal of the American Statistical Association, 1978, 73, 794-799.

Wright, B., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

SYMPOSIUM AND DISCUSSION:
STATE OF THE ART OF ADAPTIVE TESTING AND LATENT TRAIT TEST THEORY

GERHARD FISCHER
UNIVERSITY OF VIENNA

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

JAMES LUMSDEN
UNIVERSITY OF WESTERN AUSTRALIA

DAVID J. WEISS
UNIVERSITY OF MINNESOTA

COMMENT:
JOHN B. CARROLL
UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL

GERHARD FISCHER
UNIVERSITY OF VIENNA

When I first became acquainted with computerized adaptive testing, I considered it to be of little practical importance, for which psychologist is equipped with a sufficient number of computer terminals and has access to a time-sharing computer system? Therefore, I predicted only a few applications of adaptive testing in the near future. The actual development has proved me wrong. The advent of microprocessor techniques in particular is making adaptive testing practical, and adaptive testing procedures are being rapidly developed along with the spread of their applications in large-scale testing projects. Moreover, there have been advances in the theory underlying adaptive testing, e.g., the Bayesian approach. An effort to catch up with this progress will have to be made in the European countries, where, however, the number of testees is usually much smaller than in the U.S., rendering the economic aspects of adaptive testing somewhat different.

Although the theoretical advantages of adaptive testing cannot be disputed in principle, caution should be exercised against being over-enthusiastic about adaptive testing, since results from empirical applications might turn out somewhat less favorably than in theory.

Adaptive testing has become possible only through the various strong true-score theories, which--in contrast to the tautological asumptions of classical test theory--attempt to force the responses of subjects into the corset of restrictive model assumptions; the bonus from these assumptions, however, is that there is a basis for explaining observed behavior in terms of certain item and person parameters, so that the chances of a subject to solve any additional item can be predicted from previous responses, and thus an appropriate item can be chosen. The validity of this procedure, as well as that of the test results, rests wholly on the validity of the model used, and it must be required to hold for each and every subject. No non-fitting subjects, such as Lumsden´s (1980) "lazy subject" who responds inadvertently to an item, are allowed; no systematic differences between subjects or groups of subjects with respect to the ROC curves are allowed, either. Hence, the ROC curves must be the same for all subjects or, more practically, for all relevant groups of subjects within the population of interest. There will have to be a comparison of the results of item calibrations in subsets of subjects who differ as much as possible in some relevant variables, such as age, sex, socioeconomic status, education, and ability. Only if the ROC parameters come out the same in all such subgroups will the model hold with sufficient accuracy to allow adaptive testing.

The question arises, If such studies are undertaken, is there much hope for attaining stable results? To be more concrete, are the same item parameters really obtained, e.g., in groups of very bright and groups of rather dull examinees? In view of the generally acknowledged difficulties of estimating the guessing and discrimination parameters at all, it is doubtful that the estimates of these parameters would show only sampling errors when estimated from, say, groups differing radically in average latent ability. If I am correct, however, the consequence is that the validity of adaptive testing procedures based on the ROC parameters must be doubted.

Where does that leave us? Should we not resort to a model that is based on the principle that item parameter estimates must be independent of the sample of subjects, i.e., where the parameter estimates are "sample-free?" Of course, no model can guarantee what the data will be like, but the model should have a formal structure, which in principle enables the estimation of the item parameters independently of the ability distribution in the sample of subjects. In other words, this leads directly to the Rasch model.

There are some important advantages of the Rasch model with respect to adaptive testing that have not been discussed at this conference so far: By putting a linear structure into the item parameters, yielding the so-called linear logistic test model (LLTM), one can--at least in certain domains of ability testing--explain the item difficulty in terms of more elementary cognitive operations. This entails the possibility of defining large unidimensional universes of test items where each item has a difficulty parameter predicted from the logical structure of the item. The LLTM has been applied, for example, to materials similar to the Raven Progressive Matrices, however with items constructed systematically on the basis of a defined set of cognitive operations. The universe of these items is, in principle, unlimited; but in practice, of course, just a fairly large set of items is obtained. Such items have been used by Fischer and Pendl (1977) for the purpose of a simple adaptive testing strategy that can be applied without using a computer.

Besides the theoretical nicety of the LLTM, which explains item difficulty on the basis of a psychological microtheory, and besides its applicability to adaptive testing strategies, the LLTM permits an investigation of other types of problems, which could be considered as further advances in latent trait theory. It lends itself to analyzing the effects of context, item position, and learning that occurs during test-taking; to predicting the asymptotic difficulty of cognitive operations and/or of items after infinitely long practice; and, generally speaking, to analyzing the effects of any kind of experimental condition on the probability of a correct response. Furthermore, these linear logistic models have been developed and tentatively applied to polychotomous items, which yield more detailed information than the dichotomously scored items. Also, the application of the polychotomous Rasch model to projective test data, for example, has been seen to be quite successful. Going beyond the LLTM- and LLRA-type models, a dynamic extension of the Rasch model has been developed by Kempf (e.g., 1977; Kempf & Mach, 1975), viewing test-taking behavior of the subject as a stochastic process; it takes response contingent "transfer effects" on the subject's ability into account.

In conclusion, I would like to briefly discuss the progress of latent trait theory beyond the traditional domain of test theory. As was pointed out in Fisher (1980), there is an attempt to apply latent trait theory to other fields than traditional ability testing, e.g., to problems in applied and clinical psychology. One major problem type is the detection and assessment of change; that latent trait theory has been extended to multidimensional item sets is an important step. Anyone dealing with measurement of change under the influence of educational programs or therapeutical treatments will find that using unidimensional tests for measuring change means leaving out many criteria (items) that, according to the applied or clinical psychologist, are often the most relevant ones. A homogeneous test is something beautiful for the psychometrician, but it may be rather useless from the point of view of the applied psychologist. Therefore, it is a major advance that latent trait models can be adapted to multidimensional item sets, i.e., to such item sets as are approved by our colleagues from the applied departments.

Latent trait models have also been devised for analyzing types of observations that are quite different from those discussed at this conference, e.g., for describing social interaction in groups. Scheiblechner (e.g., 1977) has developed such models for qualitative observations and for frequency data. These new approaches to certain problems in social psychology seem to be quite promising. I believe, therefore, that we are at the beginning of an era of psychometrics where latent trait theory will be greatly generalized so as to become applicable to very different problems in experimental, social, and applied psychology.

REFERENCES

Fischer, G. H. Some latent trait models for measuring change in qualitative observations. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Fischer, G. H., & Pendl, P. Individualized testing on the basis of the Rasch model. Paper presented at the Third International Symposium on Educational Testing, Leyden, The Netherlands, June, 1977.

Kempf, W. F. Dynamic models for the measurement of traits in social behavior. In W. F. Kempf & B. Repp (Eds.), Mathematical models for social psychology. New York: Wiley, 1977.

Kempf, W. F. & Mach, G. A FORTRAN program for CML estimation in a dynamic test model. In W. F. Kempf, P. Hampapa, & G. Mach (Eds.), Conditional maximum likelihood (Nr. 13). Kiel: University of Kiel, Institut fur die Padagogik der naturwissenschaft, 1975.

Lumsden, J. Discussion. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Scheiblechner, H. The social structure of large groups. In W. F. Kempf & B. Repp (Eds.), Mathematical models for social psychology. New York: Wiley, 1977.

FREDERIC M. LORD
EDUCATIONAL TESTING SERVICE

I am concerned that much of the data that is gathered may be seriously impaired by students who do not cooperate, which was not a problem 20 or 30 years ago but is a serious problem in many cases today. If a student answers half the items in a normal fashion and then answers the rest of the items at random, this will create problems in the statistical analysis. If some of the students in one group in an equating study respond in this manner and those in the other group do not, the value of the study could be destroyed.

An important point is sometimes ignored in the consideration of adaptive testing: Adaptive testing is most useful when it is necessary to measure well at both extremes of the ability range. It is not at all useful if all that is needed is to divide a group of people into those who will be accepted and those who will be rejected.

I would like to endorse Lumsden´s suggestion that item parameters can be estimated much better if an extra group of low ability, and perhaps an extra group of high ability, is added to the group of subjects. If this is to be done, however, it would be very difficult to use a Bayesian approach, since it can be no longer be assumed that ability is normally distributed.

I was interested in some of Yen´s results. She studied the difference between estimated θ (ability) on two parallel tests as a function of ability level, comparing the Rasch estimates of ability with the estimates from the 3-parameter model. On the surface, the results were rather startling. The 3-parameter model yields smaller differences than the Rasch model at high ability levels; but the Rasch model yields smaller differences at low ability levels. This gives the (mistaken) impression that if estimation of the ability of high-ability level people is desired, the 3-parameter model should be used; but if estimation of the ability of low-ability level people is desired, the Rasch model should be used.

I would like to explain what I think is occurring here. The Rasch estimates of ability are based on number-correct score. A person who answers 20% of the items correctly has a standard error of measurement that is about the same as if he/she had answered 80% correctly. In the case of the 3-parameter model, where there is guessing, it is obvious that low-ability people guess frequently, which introduces random error into their scores; so it is expected that the standard error of measurement will be higher at low ability levels than at high ability levels.

Since the Rasch estimator is based on number-correct score, there is no reason for the score of a low-ability person to fluctuate wildly; thus, there is a relatively small standard error under the Rasch model. Under the 3-parameter model, if it is desired to estimate the ability of low-level people, the difficult items should not be scored but thrown away, since they just add noise to the score. To go to an extreme, the 3-parameter ability estimate for a low-ability person may be based on the person's responses to just two or three items out of the entire test. Clearly, in this extreme case, such an estimate is going to have a large standard error. Nevertheless, this is the correct way to estimate ability if there is guessing. The 3-parameter model should be used in spite of the fact that it gives this large standard error.

There is a problem correlating estimates of θ. At least in conventional testing, it is quite likely that some people will be found whose maximum likelihood estimate of ability is at $-\infty$ (In tailored testing this will be avoided if there are enough easy items in the pool). If there are a finite proportion of people with $\hat{\theta}$ of $-\infty$, it is obviously impossible to compute means and variances and correlations of $\hat{\theta}$. I do not think excluding these people is a solution; the results would depend on the vagaries of the situation--on how many people are at -50, how many at -40, and so on.

For most purposes, it really does not matter very much whether a person's ability is estimated to be -6 or -20. If it did matter, clearly we should not have given the person the test we did, we should have given him/her an easier test that would allow the accurate determination of whether he/she is at -6 or -20. That we did not give him/her such an easy test suggests that we do not care whether he/she is at -6 or -20. If this is true, then it is clearly wrong to use a numerical scale that attaches much importance to such a difference.

If a Bayesian estimation procedure is used, estimates of $-\infty$ will not be obtained. This really does not get at the basic problem, however, which is that differences at the extremes of the scale are not very important. The only way to eliminate this difficulty is to transform the scale and to use numbers that represent faithfully whatever importance is attached to the differences.

One way to do this is to transform each $\hat{\theta}$ into an estimated number-correct true score, which is a monotonic transformation. The number-correct score scale is the kind of scale that we are accustomed to using. The fact that we often work with number-correct scores suggests that this scale reflects the kinds of differences considered important.

A $\hat{\theta}$ of -6 and a $\hat{\theta}$ of -20 will both transform to a true score very close to zero. That takes care of the problem. Means and standard deviations can then be computed; and different testing procedures or different teaching procedures or different estimation procedures, or whatever it is we need to compare, can be compared on this scale.

The last point is a problem on which I am currently working, which I think is rather important: ways to correct for the bias in various quantities that are estimated by LOGIST. Bias is of particular concern when doing repeated equatings. At Educational Testing Service, Form H is equated to Form G, Form I to

Form H, Form J to Form I, Form K to Form J, and so on. Sometimes there are 12 new forms a year. If there is a small bias in each of these equatings, due to the fact that the parameter estimates are biased, the bias will accumulate over a period of time and become rather serious.

JAMES LUMSDEN
UNIVERSITY OF WESTERN AUSTRALIA

"Trotsky no doubt said many foolish things. But one wise thing he said was, 'Belief without action is death!' What do test theorists believe? How do they act? Belief without action is death. Are we all, then, test theorists, dead? Yes. And not even decent corpses enriching the earth in which we decompose. We must learn to live."

My confidence in the truth of the statement above (taken from a sermon in honor of Oscar Buros) has been shaken by events of the past few months, and particularly of the past few days. The younger test theorists seem more sensitive to problems and more willing to act than I had expected.

There are problems. The papers of this conference have consistently revealed a crisis in adaptive testing. The expensive apparatus constructed by the psychoarithmeticians has not delivered as promised. It has given, at best, mediocre results and on too many occasions results that are odd--indeed, inconceivable if the model even remotely holds.

Most of the difficulties are with the 3-parameter model, and perhaps the great arithmeticians will solve them. However, this is unlikely. There are strong theoretical grounds for the belief that there can be no satisfactory solution. What can be done about it?

The multiple-choice item can be abandoned wherever possible and completion-type items can be used. There is already available a useful range of tests that can be given in that form, for example, standard well-tested items from intelligence tests: number span (forward, back, simultaneous, successive), number series, letter series, mathematical problems, and code substitution. If a program can be found to "normalize" spelling (or if we are prepared to include spelling as part of the systematic variance), then synonyms, antonyms, and verbal analogies can be added. This is no trivial list. And it would seem highly likely that imaginative use of the flexible delivery made possible by computers will greatly enlarge the possibilities for completion items.

Abandonment of adaptive testing with multiple-choice items would thus avoid the necessity for precise estimation of item parameters. Efficient adaptive testing is only possible when discrimination over a relatively wide range of ability is required and when the discriminatory power of the items is relatively high. For all other cases conventional testing is indicated. How should the conventional test be scored? If the item characteristic curve (ICC) procedures are preferred and the uncertainties of estimation in the 3-parameter model can be tolerated, they may be used.

Lord has pointed out some troublesome end effects with the $\hat{\theta}$ metric and has suggested the "true" score, given by

$$T_i = \frac{1}{n} \sum_{g=1}^{n} p_g(\hat{\theta}) \qquad [1]$$

I cannot bring myself to call anything a true score, and I suggest that a better name for T_i is the "estimated raw score." The raw score is a good estimator of the estimated raw score, typically accounting for over 95% of the variance. For most purposes, the raw score will do everything that is needed without any need to consider very seriously the item parameters.

A more powerful alternative to adaptive testing is sequential testing, which does not seem to have been seriously treated. On the basis of a short routing test, subjects can be rejected, selected, or given further testing. With appropriate tests it should be possible to better the performance of the best conventional tests and to match that of good adaptive tests.

No one has spoken at this conference about test construction--about procedures for forming and improving item banks. This should be a matter of prime concern, for obviously no amount of arithmetic is going to overcome the problems of a badly constructed test. My preference is for factor analytic procedures. These may be used in some cases to construct a strictly unidimensional test. In others, factor analysis may be deliberately used to construct a heterogeneous test. The classical item analysis procedures may operate to exclude a precious group of items measuring an important criterion-relevant ability that is not measured by the great majority of the other items. Factor analysis gives the choice of making two tests or a single heterogeneous test.

Careful test construction with completion type items is the only way to achieve a fit to the 1-parameter Rasch model. When items are constructed according to a strict specification and tested by factor analysis, then it can be guaranteed that the slopes of the ICCs will be, at least, highly similar.

Finally, let me suggest that the proper attitude for a test theorist, indeed any theorist, is lighthearted, even playful. I notice that most test theorists are solemn. Recall the Yerkes-Dodson Law. When problems are difficult, grim determination is a disadvantage rather than a help. All theoretical advances come from analogical thinking. One should try to develop a set of analogies crammed with surplus meanings that free one from the empty mathematical formulations.

I recommend that you all start, and some finish, an elementary textbook that sets out to explain ICC theory to the most mathematically inept group, say, clinical psychologists or educators. You will find that you will be searching for clarifying examples and simple analogies to make the message comprehensible. The most important spin-off of this exercise is that you will also come to a deeper understanding and intuitive grasp of your trade.

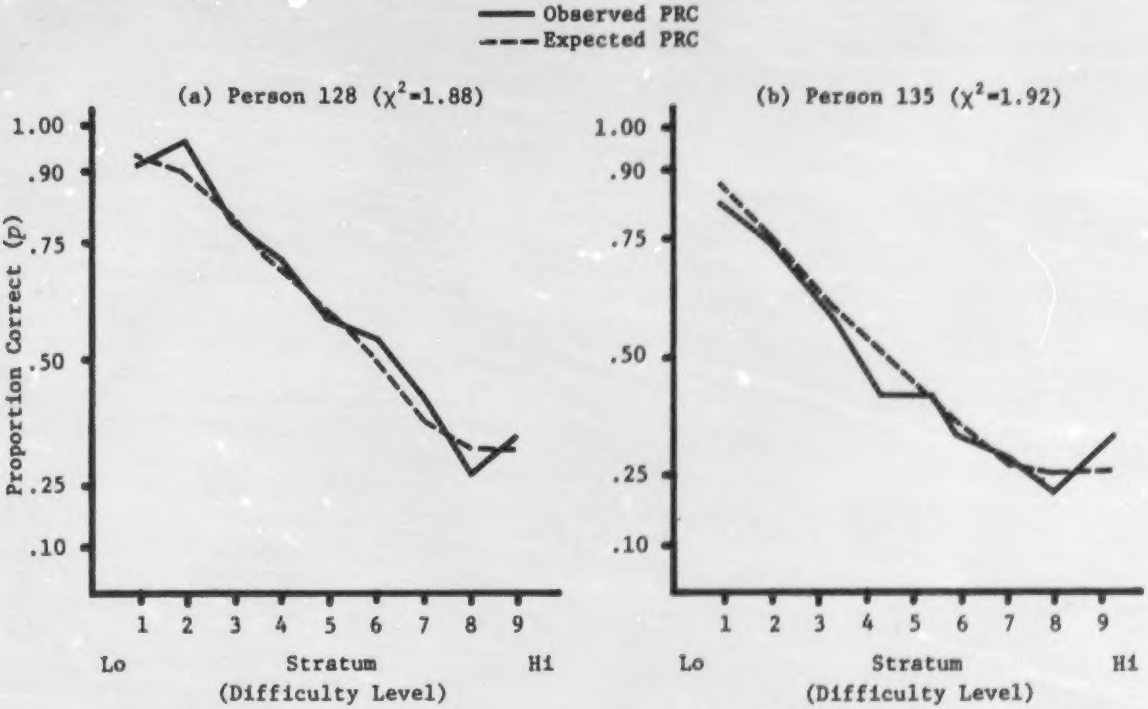
DAVID J. WEISS
UNIVERSITY OF MINNESOTA

One of the concerns that I have heard expressed during this conference has been the problem, "Do responses of real people fit the ICC model?" I began to be concerned about this problem some time ago (Weiss, 1973), resulting in my independent discovery of Mosier's (1940, 1941) Person Characteristic Curve (PCC). To investigate the idea of the PCC and to see whether it could be used to test the fit of people to the 3-parameter item characteristic curve (ICC) model, 151 students in an introductory psychology course at the University of Minnesota were administered 216 five-option multiple-choice vocabulary test items. The items were then split into subgroups by their difficulty ($\underline{b}$) parameters, and 9 strata were constructed in terms of difficulty with 24 items in each stratum. Each stratum was split into two parallel substrata. As a result, for each individual there were 18 peaked tests. Within each of those strata and substrata the proportion correct for each individual was determined. The plot of these data for one individual is an observed PCC. Several observed PCCs are shown as solid lines in Figure 1. These curves show how people differ in terms of how they obtain different proportions correct on easy items, on items of average difficulty, and on difficult items.

Given this observed data, some index was needed of whether or not the data from these students fit the model. Using the equation for the 3-parameter logistic model, the ICC parameter estimates for the items, and a maximum likelihood ability estimate for each student based on all 216 items, the estimated probability of a correct response was computed for each item. To obtain a model-predicted proportion correct for each stratum (and substratum), these estimated probabilities were summed for each stratum (and substratum) and divided by the number of items in the stratum (or substratum). These model-predicted values are shown in Figure 1 as dashed lines for each individual. Their location along the ability continuum is a function of the $\underline{b}$ values of the items and the ability estimate for the individual; the slope of the model-predicted PCC is a function of the item discriminations and guessing parameter values.

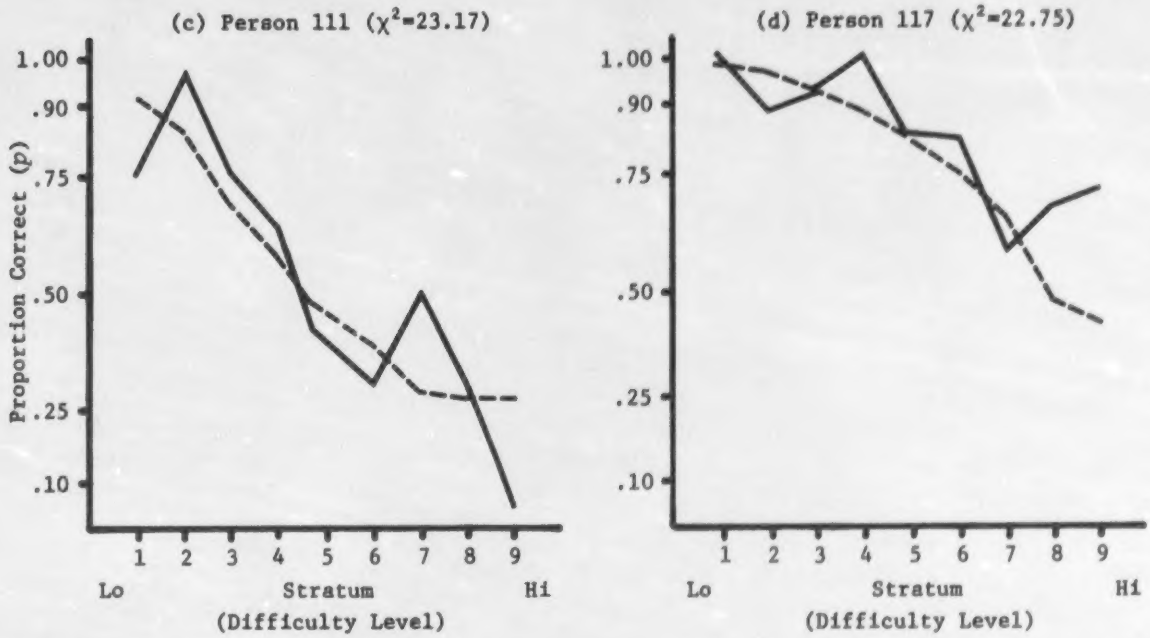
The fit of each person's observed PCC data to the model-predicted data was determined by a chi-square test. Results showed that about 90% of the students did not deviate significantly from the model at the 5% level. It was thus encouraging to see that the responses of most of the students fit the model; Figures 1a and 1b illustrate PCC data for two students whose responses did fit the model. To determine if any of that 10% group reliably did not fit the model, PCCs for each person were determined separately for each of the parallel substrata, and for each person two indices of fit were computed. Chi-square values for the first set of substrata were plotted against those for the second set,

Figure 1
Observed and Expected Person Response Curves (PRCs) for Two Persons Whose Responses Reliably Fit and Two Persons Whose Responses Did Not Reliably Fit the Three-Parameter ICC Model



(a) Person 128 (χ^2=1.88)

(b) Person 135 (χ^2=1.92)



(c) Person 111 (χ^2=23.17)

(d) Person 117 (χ^2=22.75)

and individuals whose data were significant at the 5% level for both chi-squares were identified. This analysis identified a small group of students who reliably did not fit the model. The observed and expected PCCs for two members of this group are shown in Figures 1c and 1d. These figures show two different patterns of non-fit to the 3-parameter model. But the major conclusion was that the vast majority of the students did perform in accordance with the 3-parameter model (a complete report of this study is in Trabin & Weiss, 1979).

Another theme that was prevalent at this conference was the question of whether adaptive testing should be used at all. Lumsden said it should not; Lord said it should not; Fischer said it should not; I say it should. However, we should carefully evaluate the question of fixed length versus variable length adaptive tests. Although several psychometricians supported fixed length adaptive tests, I believe that variable length tests are more appropriate than fixed length tests. This belief is based on data from the Bayesian posterior variances or the estimated standard errors of measurement for individuals taking an adaptive test; at any given item length there are individual differences in those error estimates. Some individuals are more precisely measured at a given number of items than others; and this is a function of the individuals taking tests, not a function of the item parameters themselves. It is also a function of the specific items that those individuals took. Not all items in any real item pool, regardless of how ideal it is, will be equally distant from the ability level of every person. Consequently, as long as item parameters differ in the pool, if items are selected to maximize some function for an individual, any two individuals will obtain different errors of estimate/measurement. When that happens, variable length adaptive tests are more appropriate than fixed length adaptive tests. Testing should therefore continue until the level of precision desired is obtained. Test length will then vary based on how each particular individual happens to interact with that particular subset of items; that interaction may include personality characteristics, such as risk-taking, that affect test scores but are not on the same dimension that is being measured with a particular subset of items.

A third problem that I have observed throughout this conference, mentioned earlier by Lord, which has still not been solved, is the scoring problem for latent-trait-based procedures. The Bayesian scoring procedure that is now popular has the problem of regressing ability estimates toward the mean. This means that there are some individuals whose true ability levels are two or three standard deviations away from the mean but whose ability estimates will be less extreme. The result is less discrimination among those whose ability is very high or very low using the Bayesian procedure. This problem needs to be resolved.

One solution would be a distribution-free Bayesian scoring procedure that is not a maximum likelihood procedure, since the maximum likelihood procedure has the problem of an inability to provide ability estimates for individuals with unusual response patterns (and real people do get unusual response patterns) or for individuals who answer all the items correctly or incorrectly. One ad hoc solution to the problem with the maximum likelihood estimates of ability is simply to look at the data by plotting the likelihood function for a response pattern that does not converge. This may help in uncovering the cause for the lack of convergence and will show where the likelihood function begins to flatten. This value may then be utilized as a provisional estimate of ability. This may be better for selecting subsequent test items than assigning an arbitrary -10, -40, or -5 as an ability estimate.

The c parameter in ICC theory is a problem in estimation, since it creates problems in the estimation of the a parameter and lowers estimated item discriminations. Guessing also introduces into test scores many variables that are inappropriate. Thus, I can only support Lumsden's suggestion, that the multiple-choice item be retired and that new item types be developed that are free of the technology under which testing developed 70 years ago. The new test item need not necessarily be completely free response. There are other kinds of items that will do a good job of measuring that are not necessarily free response items. Although free-response (or completion) items are obviously the ideal toward which we should strive, we should carefully examine our test items to determine whether a non-multiple-choice format can be used so as to eliminate the guessing problem and thereby do a better job in item parameter estimation and individual measurement.

At the same time we need new kinds of tests. Given the capabilities of the computer, now that we have it, we need to develop new kinds of tests that may not be based on latent trait theory but that more fully utilize the capability of the computer to interact with an individual in order to measure abilities that we are not now measuring. I hope that when we do develop these kinds of tests that we avoid the multiple-choice item and that we try to be more creative and develop testing situations that will more truly reflect the potential actual performance of people in the real world and the criteria that we are attempting to predict.

Now that we are using computers for test administration, I see a danger in the use of response latency information without carefully examining its characteristics. It is very easy now to collect response latency data on an individual taking a test item and to use those data in ways that experimental psychologists have done for many years. But there is a critical difference between what the experimental psychologist does and what the psychometrician does. The difference is that when experimental psychologists use response latency data, they typically take numerous observations and then compute mean response latencies for individuals--their means are computed either across individuals and/or over replications of stimuli--and those means average out many random fluctuations that occur in real data.

When psychometricians look at latencies for individual test items and build models about response latency for people taking individual ability test items, they might build those models on much irrelevant data. Before such models are built, the psychometrician should observe people taking an ability test. What will be observed as components of response latencies are people scratching their heads, fixing their contact lenses, observing others walking to and from their testing terminals, or just plain inattention and daydreaming, rather than responding instantaneously as soon as they have arrived at the correct answer, as the models will posit. As a result, latencies measured at the individual item level will include many random components. Elimination of these kinds of disturbances will require many replications of items with similar difficulties;

then we might obtain a valid estimate of the response latency for a person on an item subset of a given difficulty. Thus, before we attempt to use response latency data in the measurement process, the reliability and validity of reponse latency data derived from ability testing situations need to be examined.

An additional problem in adaptive testing that needs further research is the dimensionality problem. All of latent trait theory that has been studied and applied to date is based on the unidimensional case; we still have not adequately solved the multidimensional case. If latent trait theory is to be adequately used in many practical testing situations, the multidimensional case will need to be operationalized, since tests cannot always be made as unidimensional as we would like to have them.

Finally, we should not rely totally on ICC theory for adaptive testing. There are ways to implement adaptive testing that do not require ICC theory (e.g., Weiss, 1975). ICC theory will be useful if there are 1,000 subjects and 80 items (or whatever future research discovers to be adequate) on which to parameterize test items. But there are many environments where such item pools and sample sizes are not available. In these cases other ways of doing adaptive testing, which might operate more effectively than ICC methods (e.g., Thompson & Weiss, 1980), should be considered.

REFERENCES

Mosier, C. I. Psychophysics and mental test theory: Fundamental postulates and elementary theorems. Psychological Review, 1940, 47, 355-366.

Mosier, C. I. Psychophysics and mental test theory II. The constant process. Psychological Review, 1941, 48, 235-249.

Thompson, J. G., & Weiss, D. J. Criterion-related validity of adaptive testing strategies (Research Report 80-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1980. (NTIS No. AD A087595)

Trabin, T. E., & Weiss, D. J. The Person Response Curve: Fit of individuals to item characteristic curve models (Research Report 79-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1979.

Vale, C. D. & Weiss, D. J. A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, October, 1975. (NTIS No. AD A018758)

Weiss, D. J. The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, September 1973. (NTIS No. AD 768376)

JOHN B. CARROLL
UNIVERSITY OF NORTH CAROLINA
AT CHAPEL HILL

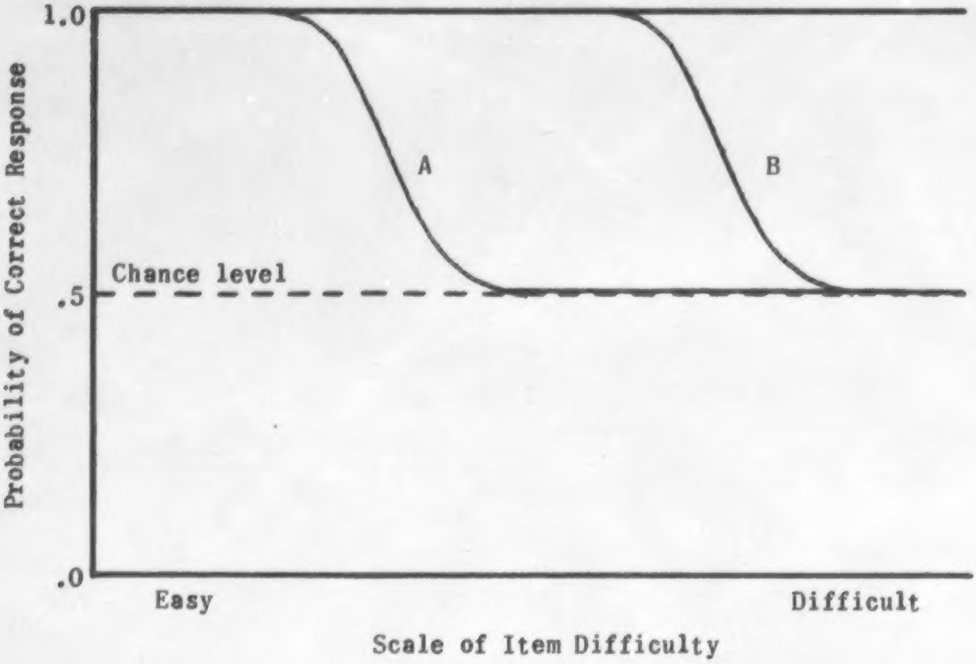
I should like to make some comments about the person characteristic curve (PCC), which has just been discussed by Weiss (1980). Before moving into the field of cognitive psychology, I was a test theorist; and one of my concerns, although not under that name, was actually the PCC. My original interest stemmed from a paper by Guilford (1941) in which he claimed that a factor analysis of the 10 subtests of the Seashore Test of Pitch Discrimination revealed that more than one ability would be involved in performance on this test. Indeed, he believed that three factors were involved--one for easy items, one for items of medium difficulty, and one for difficult items.

This conclusion made absolutely no sense: I found it difficult to believe that an individual who could not make an easy pitch dicrimination could nevertheless detect a very small pitch difference. I developed the statistical rationale (Carroll, 1945) whereby I was able to convince myself that Guilford's findings were an artifact resulting from the use of tetrachoric correlations with the scores affected by chance success--a conclusion that Gourlay (1951) confirmed and that I have discussed (see Carroll, 1961).

Underlying this rationale was the notion that the response curve of an individual to items of varying difficulty measuring a single trait was a psychometric function, for example, a normal ogive starting at probability asymptotic to unity for easy items and descending to near zero, or at least to a chance level $\underline{c}$, for more difficult items. Actually, this is simply a version of a standard psychophysical function. It is well illustrated with data from the Seashore Test of Pitch Discrimination, which contains 10 subtests, each with 10 two-choice items at a particular level of difficulty in terms of the difference (in Hertz) between the two pitches presented for a judgment as to whether the second pitch is higher or lower than the first. Because of unreliability and chance success factors, the response curve for any single individual will be rather irregular; but mean response curves for individuals at different total test score intervals will exhibit the form illustrated in Figure 1. In effect, these are mean PCCs and they are similar to Weiss's (1980) illustration for a vocabulary test.

I, too, have plotted such curves for vocabulary tests, as well as for achievement test items, as in a study of Navy officer candidate examinations (Carroll & Schohan, 1953), where they were called individual operating characteristic curves. I have tended to think of the slopes of these curves as indicating something about the trait being measured, rather than the individual. With a psychophysical function such as that of pitch discrimination, the slopes

Figure 1
Expected Mean Person Response Characteristic Curves
for (A) Low-Ability Examinees and (B) High-Ability Examinees
on a Test of a Trait such as Pitch Discrimination Ability



will be relatively steep; but with achievement tests, the slopes will be relatively less steep. In fact, in the case of the Navy officer candidate examinations (Carroll & Schohan, 1953), the slopes were so low as to indicate that the tests were "perfectly heterogeneous tests"; that is, they were the slopes that would be expected for tests composed of items differing in difficulty but with population intercorrelations (corrected for chance success effects) equal to zero. Even though my other interests and commitments have never permitted me to develop this kind of test theory as much as I would have wished, I recommend that this line of thinking be further explored, particularly in the light of latent trait theory. (An exposition and application of my theory, as far as I carried it, is to be found in a doctoral thesis by Dry, 1959.)

One interesting point emerges. Contrary to some opinions that have been expressed here—opinions that can be respected in view of the reasons given for them—I am going to be very heretical and suggest that rather than "getting rid of" multiple-choice items, we feature them in our work but make them two-choice instead of "multiple" choice. This is essentially what many experimental cognitive psychologists have been doing: converting multiple-choice tests to a two-choice format in order to capitalize on certain advantages of this format. For example, Egan (1976) converted several of Guilford's spatial ability tests to a two-choice format, primarily in order to obtain more valid and reliable response latency measurements. Giving the respondent a two-choice option (a true-false or a yes-no option) obviates the problem of time wasted in scanning, comparing, and evaluating a large number of choices. This is one advantage of the two-choice format. Another advantage, from the standpoint of latent trait theory, is that the *c* parameter can be determined, in many circumstances, by a priori considerations as equal to .5, provided that the examinee is led to believe that the probability of a particular response being correct is .5. This can be done, of course, by insuring that equal numbers of true-false (or yes-no) items are present in the test or the experimental series.

Note that experimental psychologists are not usually interested in the subject's latency or correctness on a single item; they take measurements over groups of similar items or replicate the data over multiple trials or trial blocks. A similar approach can be taken in the case of ability or achievement testing without increasing testing time much, if at all. Actually, constructing large numbers of two-choice items is easier than constructing large numbers of five-choice items. However, one should avoid making items that deliberately mislead low-ability examinees into making incorrect responses, for in this case the *c* parameter can easily drop well below .5.

It has been my intention in this discussion to mention some possibilities that might well be followed up in future work on the applications of latent trait theory to computerized testing; I will be interested in watching any such developments.

REFERENCES

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 1945, *10*, 1-19.

Carroll, J. B. The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 1961, *26*, 347-372.

Carroll, J. B., & Schohan, B. *Construction of comphrehensive achievement examinations for Navy officer candidate programs* (Project NR 154-138). Pittsburgh, PA: American Institute for Research, November 1953.

Dry, R. J. *An application of a theory of pure factor tests to the construction of homogeneous tests*. Unpublished doctoral dissertation, Harvard University Graduate School of Education, 1959.

Egan, D. E. *Accuracy and latency scores as measures of spatial information processing* (Research Report No. 12224). Pensacola, FLA: Naval Aerospace Medical Research Laboratory, February 1976.

Gourlay, N. Difficulty factors arising from the use of tetrachoric correlations in factor analysis. *British Journal of Statistical Psychology*, 1951, *4*, 65-73.

Guilford, J. P. The difficulty of a test and its factor composition. Psychometrika, 1941, 6, 67-77.

Weiss, D. J. Discussion. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

ADDRESSES OF CONFERENCE PARTICIPANTS*

Tom Beisaker
Data Design Laboratories
7925 Center Avenue
Cucamonga, CA 91730

James Blacksher
Personnel and Training
Evaluation Program
Naval Guided Missile School
Damneck
Virginia Beach, VA 23461

R. Darrell Bock
Department of Education
University of Chicago
5835 Kimbark Road
Chicago, IL 60637

Commander Arnold Bohrer
Centrum Recrutering en Selectie
Kazerne Klein Kasteeltze
1000 Brussels, BELGIUM

Robert Brennan
ACT
P.O. Box 168
Iowa City, IA 52240

Elana Broch
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Jerry Buchmeier
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Wolfgang Buchtala
Fa. Ing. Bruno ZAK GmbH
Postfach 1306
Industriestrasse 1
D-8346 Simbach/Inn, WEST GERMANY

John B. Carroll
Psychometric Laboratory
Davie Hall 031A
University of North Carolina
Chapel Hill, NC 27514

Austin T. Church
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Leslie Crichton
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Joe E. Crick
Computer Center
Harbor Campus
University of Massachusetts
Boston, MA 02125

Charles E. Davis
Office of Naval Research
536 South Clark Street
Chicago, IL 60605

Charles Dunbar
ACT
P.O. Box 168
Iowa City, IA 52243

Marshall J. Farr
Director, Personnel and Training
Research Programs
Office of Naval
Research (Code 458)
Arlington, VA 22217

Gerhard Fischer
Psychlogisches Institut
der Universitaet Wien
Liebigasse 5
A-1010 Vienna, AUSTRIA

Hideo Fujiwara
Faculty of Engineering
Department of Electronic Engineering
Osaka University
Suita,
Osaka 565, JAPAN

Kathleen A. Gialluca
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Janice Gifford
School of Education
University of Massachusetts
Amherst, MA 01003

Steven Gorman
Assistant to the Deputy
Chief of Naval Operations
(Manpower, Personnel and Training)
OP-01T Naval Bureau of Personnel
Arlington Annex, Room 2705
Washington, DC 20370

William Graham
Testing Directorate
Military Enlistment
Processing Command
MEPCT-P
Fort Sheridan, IL 60037

Bert F. Green, Jr.
Department of Psychology
Johns Hopkins University
Baltimore, MD 21218

Ronald K. Hambleton
School of Education
University of Massachusetts
Amherst, MA 01002

Lutz Hornke
Universitaet Duesseldorf
Erziehungswissenschaftliches
Institut
Universitaetsstrasse 1
D-4000 Duesseldorf, WEST GERMANY

Gail Ironson
Psychology Department
107 Social Science Building
University of South Florida
Tampa, FL 33620

Carl Jensema
Office of Demographic Studies
Gallaudet College
Kendall Green
Washington, DC 20002

Harold Jensen
Air Force Human Resources Laboratory,
Personnel Research Division
Brooks Air Force Base
San Antonio, TX 78235 (AFSC)

Marilyn Johnson
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Douglas Jones
Educational Testing Service
Princeton, NJ 08541

Stanley Kalisch
Educational Testing Services
Suite 1040
3445 Peachtree Road NE
Atlanta, GA 30326

G. Gage Kingsbury
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Neal Kingston
Educational Testing Service
Princeton, NJ 08541

Charles Kreitzberg
Educational Testing Service
Princeton, NJ 08540

Guergen Kulling
Streitkraefteamt
D-5300 Bonn 2, WEST GERMANY

George W. Lawton
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Michael Levine
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

Philip Livingston
Department of Psychology
University of Tennessee
Knoxville, TN 37916

Frederick M. Lord
Educational Testing Service
Princeton, NJ 08540

James Lumsden
Department of Psychology
University of Western Australia
Nedlands 6009, AUSTRALIA

James McBride
Code P310
Navy Personnel Research
and Development Center
San Diego, CA 92152

---

*Addresses current as of October 1980.

John T. Martin
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Vincent Maurelli
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Ronald J. Mead
1925 Dupont Avenue South
Minneapolis, MN 55403

Phillip Metres
Office of Naval Research
536 South Clark Street
Chicago, IL 60605

Anne Morgan
Bureau of Social Science Research
1990 M Street, N.W
Washington, DC 20036

Melvin R. Novick
Division of Educational Psychology,
Measurement and Statistics
356 Lindquist Center
University of Iowa
Iowa City, IA 52242

Mario Padron
Department of Psychology
University of Illinois
Champaign, IL 61820

Wayne Patience
Department of Educational Psychology
4 Hill Hall
University of Missouri
Columbia, MO 65201

Michael Patrow
Headquarters Marine Corps
(Code MPI-20)
Washington, DC 20380

James A. Paulson
Portland State University
P.O. Box 751
Portland, OR 97207

Nancy Petersen
Educational Testing Service
Princeton, NJ 08540

Frank Petho
Naval Aerospace Medical Research
Laboratory (Code 15)
Pensacola, FL 32508

Steven Pine
4950 Douglas Avenue
Golden Valley, MN 55416

J. Stephen Prestwood
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Mark Reckase
Department of Educational Psychology
University of Missouri
4 Hill Hall
Columbia, MO 65201

Malcolm James Ree
Air Force Human Resources Laboratory,
Personnel Research Division
Brooks Air Force Base
San Antonio, TX 78235 (AFSC)

Robert Ross
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Fumiko Samejima
Department of Psychology
University of Tennessee
Knoxville, TN 37916

Lalitha Sanathanan
Argonne National Laboratory
Environmental Impact Studies
Division
9700 S. Cass Avenue
Argonne, IL 60439

Michael Sauter
Universitaet Duesseldorf
Ersiehungswissenschaftliches
Institut
Universitaetsstrasse 1
D-4000 Duesseldorf, WEST GERMANY

Lynn Schwendeman
Data Design Laboratories
Kitsap Center
P.O. Box 952
Silverdale, WA 98383

Major Paul Sieben
SGE/SERS
Quartier Reine Elisabeth
Rue d'Evere
7740 Brussels, BELGIUM

Rick Steinheiser
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Martha Stocking
Educational Testing Service
Princeton, NJ 08541

Deb Suhadolnik
Department of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis, MN 55455

Hariharan Swaminathan
Department of Education
University of Massachusetts
Amherst, MA 01003

Leonard Swanson
Educational Testing Service
Princeton, NJ 08541

James B. Sympson
Educational Testing Service
Princeton, NJ 08540

Kikumi Tatsuoka
Computer-Based Education Research
Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

Maurice Tatsuoka
Department of Educational Psychology
University of Illinois
Urbana, IL 61801

David Thissen
Department of Psychology
426 Fraser Hall
The University of Kansas
Lawrence, KS 66045

Robert L. Trestman
Department of Psychology
University of Tennessee
Knoxville, TN 37916

Robert Tsutakawa
Department of Statistics
224A Math Sciences Building
University of Missouri
Columbia, MO 65211

C. David Vale
Assessment Systems Corporation
2395 University Ave.
Suite 306
St. Paul, MN 55114

Howard Wainer
Educational Testing Service
Princeton, NJ 08540

Michael Waller
Department of Educational Psychology
Enderis Hall
University of Wisconsin
Milwaukee, WI 53201

Thomas Warm
U.S. Coast Guard Institute
P.O. Substation 18
Oklahoma City, OK 73169

Brian Waters
Human Resources Research Organization
300 N. Washington Street
Alexandria, VA 22314

David J. Weiss
Department of Psychology
N660 Elliott Hall
75 E. River Road
University of Minnesota
Minneapolis, MN 55455

Ron Weitzman
Code 54WZ
Naval Postgraduate School
Monterey, CA 93940

John Welsh
HQUSAF/MPCYPT
Personnel Testing
Randolph Air Force Base, TX 78148

Wolfgang Wildgrube
Streitkraefteamt
BOX 20 50 03
D-5300 Bonn 2, WEST GERMANY

Dennis Wilkinson
Data Design Laboratories
7925 Center Avenue
Cucamonga, CA 91730

Yoneo Yamamoto
Computer-Based Education Research
Laboratory
252 Engineering Research Laboratory
University of Illinois
Urbana, IL 61801

Wendy Yen
CTB/McGraw Hill
Del Monte Research Park
Monterey, CA 93940

DISTRIBUTION LIST

Navy

1 Dr. Robert Breaux
Code N-711
NAVTRAEQUIPCEN
Orlando, FL 32813

1 Dr. Richard Elster
Department of Administrative Sciences
Naval Postgraduate School
Monterey, CA 93940

1 DR. PAT FEDERICO
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Mr. Paul Foley
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego, CA 92152

1 Dr. Henry M. Halff
Department of Psychology,C-009
University of California at San Diego
La Jolla, CA 92093

1 CDR Robert S. Kennedy
Head, Human Performance Sciences
Naval Aerospace Medical Research Lab
Box 29407
New Orleans, LA 70189

1 Dr. Norman J. Kerr
Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington, TN 38054

1 Dr. William L. Maloy
Principal Civilian Advisor for
Education and Training
Naval Training Command, Code 00A
Pensacola, FL 32508

1 Dr. Kneale Marshall
Scientific Advisor to DCNO(MPT)
OP01T
Washington DC 20370

1 CAPT Richard L. Martin, USN
Prospective Commanding Officer
USS Carl Vinson (CVN-70)
Newport News Shipbuilding and Drydock Co
Newport News, VA 23607

1 Dr William Montague
Navy Personnel R&D Center
San Diego, CA 92152

1 CAPT Paul Nelson, USN
Chief, Medical Service Corps
Bureau of Medicine & Surgery (MED-23)
U. S. Department of the Navy
Washington, DC 20372

1 Ted M. I. Yellen
Technical Information Office, Code 201
NAVY PERSONNEL R&D CENTER
SAN DIEGO, CA 92152

1 Library, Code P201L
Navy Personnel R&D Center
San Diego, CA 92152

1 Psychologist
ONR Branch Office
Bldg 114, Section D
666 Summer Street
Boston, MA 02210

1 Psychologist
ONR Branch Office
536 S. Clark Street
Chicago, IL 60605

1 Office of Naval Research
Code 437
800 N. Quincy SStreet
Arlington, VA 22217

1 Psychological Sciences Division
Code 450
Office of Naval Research
Arlington, VA 22217

5 Personnel & Training Research Programs
(Code 458)
Office of Naval Research
Arlington, VA 22217

1 Psychologist
ONR Branch Office
1030 East Green Street
Pasadena, CA 91101

1 Office of the Chief of Naval Operations
Research Development & Studies Branch
(OP-115)
Washington, DC 20350

1 The Principal Deputy Assistant
Secretary of the Navy (MRA&L)
4E780, The Pentagon
Washington, DC 20350

1 Dr. Bernard Rimland (03B)
Navy Personnel R&D Center
San Diego, CA 92152

1 Mr. Arnold Rubenstein
Naval Personnel Support Technology
Naval Material Command (08T244)
Room 1044, Crystal Plaza #5
2221 Jefferson Davis Highway
Arlington, VA 20360

1 Dr. Worth Scanland
Chief of Naval Education and Training
Code N-5
NAS, Pensacola, FL 32508

1 Dr. Robert G. Smith
Office of Chief of Naval Operations
OP-987H
Washington, DC 20350

1 Dr. Alfred F. Smode
Training Analysis & Evaluation Group
(TAEG)
Dept. of the Navy
Orlando, FL 32813

1 Dr. Richard Sorensen
Navy Personnel R&D Center
San Diego, CA 92152

1 DR. MARTIN F. WISKOFF
NAVY PERSONNEL R& D CENTER
SAN DIEGO, CA 92152

1 Mr John H. Wolfe
Code P310
U. S. Navy Personnel Research and
Development Center
San Diego, CA 92152

Army

1 Technical Director
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Myron Fischl
U.S. Army Research Institute for the
Social and Behavioral Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Milton S. Katz
Training Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Harold F. O'Neil, Jr.
Attn: PERI-OK
Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Robert Sasmor
U. S. Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

1 Dr. Joyce Shields
ARI
5001 Eisenhower Blvd
Alexandria, VA 22333

1 Dr. Joseph Zeidner
ARI
5001 Eisenhower Blvd
Alexandria, VA 22333

Air Force

1 AFA/RRE
USAF Academy, CO 80840

12 AFHRL/TSR
Brooks AFB, TX 78235

1 AFIT
Wright-Patterson AFB, OH 45433

1 AFMPC/MPCYPR
Randolph AFB, TX 78148

1 AFOSR
Bolling AFB, DC 20332

1 Air Force Academy Library
USAF Academy, CO 80840

1 Air Force Human Resources Lab
AFHRL/MPD
Brooks AFB, TX 78235

1 Dr. Earl A. Alluisi
HQ, AFHRL (AFSC)
Brooks AFB, TX 78235

1 Dr. Genevieve Haddad
Program Manager
Life Sciences Directorate
AFOSR
Bolling AFB, DC 20332

1 Air University Library
Maxwell AFB, AL 36112

1 HQ AFROTC/Mr. Gordon
Maxwell AFB, AL 36112

1 HQ ATC/XPTIA
Randolph AFB, TX 78148

1 HQ AFSC/DLS
Andrews AFB, MD 20334

1 HQ USAF/MPX
Washington, DC 20330

1 Dr. Marty Rockway
Technical Director
AFHRL(OT)
Williams AFB, AZ 58224

1 Jack A. Thorp, Maj., USAF
Life Sciences Directorate
AFOSR
Bolling AFB, DC 20332

1 Dr. Joe Ward, Jr.
AFHRL/MPMD
Brooks AFB, TX 78235

1 USAFOMC/OMY
Randolph AFB, TX 78148

1 USAFSAM/EDK (Dr Rayman)
Brooks AFB, TX 78235

Marines

1 Director, Office of Manpower Utilization
HQ, Marine Corps (MPU)
BCB, Bldg. 2009
Quantico, VA 22134

1 Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

1 DR. A.L. SLAFKOSKY
SCIENTIFIC ADVISOR (CODE RD-1)
HQ, U.S. MARINE CORPS
WASHINGTON, DC 20380

Other DoD

1 Cdr. Paul Chatelier
OUSDR&E
Pentagon
Washington, DC 20301

12 Defense Technical Information Center
Cameron Station, Bldg 5
Alexandria, VA 22314
Attn: TC

1 Dr. Craig I. Fields
Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

1 Dr. Dexter Fletcher
ADVANCED RESEARCH PROJECTS AGENCY
1400 WILSON BLVD.
ARLINGTON, VA 22209

1 Director, Research and Data
OASD(MRA&L)
3B919, The Pentagon
Washington, DC 20301

1 MAJOR Wayne Sellman, USAF
Office of the Assistant Secretary
of Defense (MRA&L)
3B930 The Pentagon
Washington, DC 20301

Civil Govt

1 Dr. Susan Chipman
Learning and Development
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Dr. Joseph I. Lipson
SEDR W-638
National Science Foundation
Washington, DC 20550

1 Dr. John Mays
National Institute of Education
1200 19th Street NW
Washington, DC 20208

1 Personnel R&D Center
Office of Personnel Managment
1900 E Street NW
Washington, DC 20415

1 Dr. Vern W. Urry
Personnel R&D Center
Office of Personnel Management
1900 E Street NW
Washington, DC 20415

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington, DC 20550

Non Govt

1 Dr. Isaac Bejar
Educational Testing Service
Princeton, NJ 08450

1 Dr. Werner Birke
DezWPs im Streitkraefteamt
Postfach 20 50 03
D-5300 Bonn 2
WEST GERMANY

1 Dr. Nicholas A. Bond
Dept. of Psychology
Sacramento State College
600 Jay Street
Sacramento, CA 95819

1 Dr. Robert Brennan
American College Testing Programs
P. O. Box 168
Iowa City, IA 52240

1 Dr. Norman Cliff
Dept. of Psychology
Univ. of So. California
University Park
Los Angeles, CA 90007

1 Dr. Meredith P. Crawford
American Psychological Association
1200 17th Street, N.W.
Washington, DC 20036

1 Dr. Hans Crombag
Education Research Center
University of Leyden
Boerhaavelaan 2
2334 EN Leyden
The NETHERLANDS

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda, MD 20014

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City, IA 52240

1 DR. ROBERT GLASER
LRDC
UNIVERSITY OF PITTSBURGH
3939 O'HARA STREET
PITTSBURGH, PA 15213

1 Dr. Lloyd Humphreys
Department of Psychology
University of Illinois
Champaign, IL 61820

1 Dr. Huynh Huynh
College of Education
University of South Carolina
Columbia, SC 29208

1 Dr. Charles Lewis
Faculteit Sociale Wetenschappen
Rijksuniversiteit Groningen
Oude Boteringestraat
Groningen
NETHERLANDS

1 Dr. Robert Linn
College of Education
University of Illinois
Urbana, IL 61801

1 Dr. Jesse Orlansky
Institute for Defense Analyses
400 Army Navy Drive
Arlington, VA 22202

1 MINRAT M. L. RAUCH
P II 4
BUNDESMINISTERIUM DER VERTEIDIGUNG
POSTFACH 1328
D-53 BONN 1, GERMANY

1 Dr. Ernst Z. Rothkopf
Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974

1 Dr. Lawrence Rudner
403 Elm Avenue
Takoma Park, MD 20012

1 DR. SUSAN E. WHITELY
PSYCHOLOGY DEPARTMENT
UNIVERSITY OF KANSAS
LAWRENCE, KANSAS 66044